Refining the distributional hypothesis: A role for time and context in semantic representation

Melody Dye (meldye@indiana.edu)[1], Michael N. Jones (jonesmn@indiana.edu)[1], Daniel Yarlett (daniel.yarlett@gmail.com)[2], & Michael Ramscar (michael.ramscar@uni-tuebingen.de)[3]
[1] Department of Psychological & Brain Sciences, Indiana University, Bloomington
[2] Referral Exchange, San Francisco [3] Department of Linguistics, University of Tübingen

Abstract

Distributional models of semantics assume that the meaning of a given word is a function of the contexts in which it occurs. In line with this, prior research suggests that a word's semantic representation can be manipulated – pushed toward a target meaning, for example – by situating that word in distributional contexts frequented by the target. Left open to question is the role that order plays in the distributional construction of meaning. Learning occurs in time, and it can produce asymmetric outcomes depending on the order in which information is presented. Discriminative learning models predict that systematically manipulating a word's preceding context should more strongly influence its meaning than should varying what follows. We find support for this hypothesis in three experiments in which we manipulated subjects' contextual experience with novel and marginally familiar words, while varying the locus of manipulation.

Keywords: distributional semantics; vector space models; discriminative learning; word frequency; semantic priming

Introduction

In the study of human conceptual knowledge, a central theoretical question concerns how semantic representations are learned from the environment. How do speakers acquire knowledge of the meaning of a word and the precise contexts of its use? How are they able to make principled inferences about its senses and its similarity to other words? Inquiries in this domain have focused on two types of converging information sources that are thought to underpin these representations – perceptual and distributional (Andrews, Vigliocco, & Vinson, 2009; Bruni, Tran, & Baroni, 2014). ***Perceptual*** data derives from experiencing words in relation to the world, in connection with objects, events, and affordances in the immediate physical environment. ***Distributional*** data, by contrast, derives from experiencing words in relation to other words. While it is clear that neither data stream alone suffices to explain semantic representation, there appears to be considerable redundancy between them (Louwerse, 2007; Riordan & Jones, 2010).

Distributional models operate on the assumption that the similarity between two words is a function of the overlap between the contexts in which they occur, a principle commonly known as the ***distributional hypothesis*** (Firth, 1957; Miller & Charles, 1991). For instance, encountering the word *violin* in the same context as *classical* and *strings* supports the inference that these words are semantic neighbors. Such an inference will also be supported for words that occur in closely related musical contexts, such as *cello*, but not for those that occur in unrelated contexts, such as *jaguar*. One of the key advantages of the distributional approach is that it provides an objective and replicable method of quantifying meaning, based solely on the statistical regularities found in large bodies of text.

Since the introduction of Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) to the cognitive sciences, a variety of different distributional models have been proposed to account for semantic phenomena. Within this class of models, there is considerable variation in implementation (for the latest class, see Baroni, Dinu & Kruszewksi, 2014). Nevertheless, they share the same core architectural assumption that word meaning is derivable from lexical co-occurrence patterns. Words are represented as vectors within a high-dimensional semantic space, and word meanings as points located within that space. Whereas distributionally similar words tend to cluster together, words that occur in more distinctive contexts are more dispersed. The similarity relations derived from these models can then be used to account for phenomena as diverse as semantic priming (Jones, Kintsch, & Mewhort, 2006), semantic categorization (Bullinaria & Levy, 2007), and visual search (Huettig, Quinlan, McDonald, & Altmann, 2006).

Implicit in these models is the notion that the lexicon is a highly interconnected system. The representation of a given word is neither static nor modular, but changes as a function of lingustic experience, both with that word in particular, and with others within the lexicon. As a demonstration of this principle, McDonald and Ramscar (2001) manipulated readers' semantic representations of marginally familiar and novel words by situating them in paragraph contexts that also contained close associates of a target meaning. For instance, subjects who read about a *samovar* in paragraph containing words like *boil* and *electric* rated it as closer to the meaning of *kettle* than subjects who read a modified version of the paragraph, which contained associates of an alternative meaning, *urn*. Even though subjects never directly observed the word *kettle* in training, their representation of the critical item—*samovar*—was moved closer to it, simply by virtue of encountering *samovar* in a similar linguistic context.

Learning in Time

One question that arises from this, is the extent to which distributional learning about a particular item is influenced by the ***sequential structure*** of the context in which it is embedded (Elman, 1990; Jones & Mewhort, 2007). Language unfolds in time, with one word following another in succession. Thus, the influence that the local context exerts on the critical item might depend on whether it helps predict the occurrence of that item, or is, in turn, predicted

by it – that is, whether the context is encountered *before* or *after* the item.

This framing maps naturally onto the the ***convergent*** and ***divergent learning hierarchies*** described by Osgood (1949). These abstract schemas capture asymmetries in how information is structured in time (**Figure 1**). In associative learning, convergent hierarchies label a situation in which a variety of cues are associated with a functionally identical outcomes (C_1, C_2, ...$C_x \Rightarrow O$), while divergent hierarchies label the inverse scenario, in which a single cue is associated with varied outcomes ($C \Rightarrow O_1$, O_2, …O_x). Convergent hierarchies have been found to result in greater facilitation and positive transfer in learning, whereas divergent hierarchies yield interference and negative transfer.

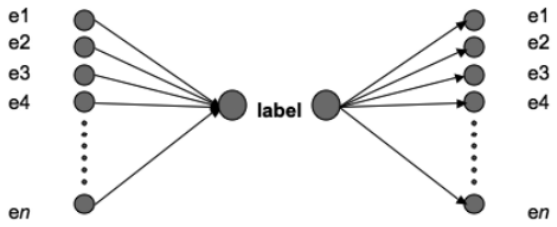

Figure 1: Sequential relationships between linguistic regularities. The left side of the figure shows a convergent hierarchy; the right, a divergent one (Ramscar, 2013).

The temporal asymmetries captured by these schemas appear to be ubiquitous in ***word learning*** (Ramscar et al., 2010). Consider the problem of learning the relation between a class of things in the world – say, the category [cat] – and the word that denotes it – *cat*. Clearly, a sizable discrepancy exists between the rich array of perceptual features that belong to the class and the comparatively sparse acoustic features of the verbal label. Whereas the flesh and blood exemplars of the category exhibit a wide variety of discriminable features, across various perceptual modalities, the label itself comprises a simple sequence of sounds, which are likely to be perceived categorically (Kuhl, 2000). Accordingly, in a standard category learning paradigm, in which a category exemplar precedes its verbal label, a convergent hierarchy results. However, simply reverse the timing—by placing the label *before* the exemplar —and the structure becomes divergent.

The terminology used to describe this pair of temporal structures varies by research domain. In the study of ***categorization***, a distinction is commonly drawn between classification, in which subjects predict the class to which an exemplar belongs based on its features, in line with a convergent schema, and inference, in which subjects predict an exemplar's feature values based on its class, in line with a divergent schema (Yamauchi & Markman, 1988). Likewise, in the study of ***causal reasoning***, predictive reasoning licenses inferences from a variety of possible causes to a shared effect—both *rain* and *sprinklers* make grass *wet*—in line with a convergent schema, whereas diagnostic reasoning licenses inferences from a common cause to its possible effects—*rain* makes grass *wet* and *green*—in line with a divergent schema (Waldmann & Holyoak, 1992; Waldmann, 2000).

Learning algorithms can help provide a mechanistic account of how the structure of information in time affects what is learned in these tasks. A critical assumption shared by most models of learning, ranging from classical conditioning to perceptrons, is that learning is scaffolded by the ***predictions*** we make about our environments, and powered by the ***surprise*** we experience whenever there is a mismatch between expectation and reality. Learning proceeds as a continual process of updating and refining expectations, selectively weighting the most informative predictors to relevant outcomes, while eliminating redundant or potentially misleading cues. When our predictions align with reality, learning asymptotes (Rescorla, 1988).

To examine how convergent and divergent structures affect word learning, Ramscar and colleagues (2010) simulated supervised category learning with the Rescorla-Wagner rule, while manipulating the sequencing of category exemplars and verbal labels. The findings were striking: The same algorithm, run over the same task, produced remarkably different representations of the learning environment, depending on the temporal sequencing of information: While convergent structures yielded ***predictive*** representations, divergent structures yielded ***veridical*** ones. Specifically, convergent schemas facilitated competition between the available perceptual features for associative weight, resulting in abstraction of the informative dimensions that best predicted the category label. By contrast, divergent schemas facilitated learning of the actual feature probabilities given the label. (For a closely related result in a different model architecture, see Yamauchi, Love, & Markman, 2002).

The differences in these representations can be mapped onto the differences between ***discriminative*** and ***generative classifiers*** in machine learning (Ng & Jordan, 2002). In learning a verbal category, the problem is to establish the likelihood of a category label L given some set of perceptual features F. To solve this problem, discriminative classifiers learn a direct mapping between features and labels, which yields $p(L|F)$. Generative classifiers solve the same problem indirectly, by learning the joint probability of $p(L, F)$ and relying on Bayesian inference to calculate the posterior likelihood of $p(L|F)$. While discriminative classifiers are more efficient and better at minimizing error, generative classifiers operate with a more complete picture of the probability space (Levering & Kurtz, 2014). Convergent schemas yield $p(L|F)$; divergent schemas $p(L, F)$.

The resultant representations appear to be optimized for different tasks. In studies of human category learning, convergent schemas benefit the learning of categories that require information-integration (Ashby, Maddox, & Bohil, 2003; Yamauchi et al., 2002), which likely form the majority of natural kinds (Rosch & Mervis, 1975). However, there are notable drawbacks to categorical responding. As a category structure becomes better learned, stimulus dimensions that are relevant to a particular categorization are selectively attended, such that they acquire distinctiveness, while irrelevant dimensions are

ignored, or down-weighted, maximizing intra-category similarities and inter-category differences (Goldstone, 1994; Lawrence, 1949; Nosofsky, 1986). Accordingly, while convergent schemas support accurate categorization across an array of perceptual domains, they can also systematically alter similarity relations, impairing memory for exemplars seen in training (Davis & Love, 2010; Dye & Ramscar, 2009) and distorting judgements of the underlying featural space (Yamauchi & Markman, 1998). Likewise, in causal inference, whereas predictive reasoning is susceptible to blocking effects, diagnostic reasoning is not (Waldmann & Holyoak, 1992). The optimal information structure at encoding thus depends on the demands imposed at retrieval (Tulving & Thomson, 1973).

Previous research has examined the effect of these asymmetric information structures on category learning and causal inference. This paper addresses itself to distributional learning, where what is learned is not the relation between words and physical referents, but rather that of words in relation to each other.

Study	**1**	**2**	**3**
Cover Story	Alien Grammar	Man vs. Machine	Semantic Identification
Training Design	10 Train-Test Blocks	1 Train-Test Block	1 Train-Test Block
Training Length	8 Associates / Topic	15 / Topic	15 / Topic
Critical Item	Pseudoword	LF	LF & HF
Topic Meanings	Random Assignment	Synonyms of Critical Item	Semantic Category

Table 1: Design differences between studies.

Studies

In the following three experiments, our aims were first, to build on the original findings of McDonald and Ramscar (2001)—which demonstrated that a pair of words can be moved closer together in semantic space even if they have never been encountered together—and second, to investigate whether readers would attach more weight to the associates that occurred *before* a word of interest, rather than *after*, as predicted by previous simulations (Ramscar et al. 2010).

The three studies presented here are all variations on the same principal design. In training, subjects read short passages containing critical words. These passages had been constructed such that the contexts occurring *before* the critical item were designed to encourage one set of inferences about its meaning, while the contexts occurring *after* it were designed to encourage a different, competing set of inferences. This design allowed us to measure the relative influence of preceding and succeeding contexts on semantic representation.

Variations on this design were devised to investigate the robustness of the predicted effects of training, and included (e.g.) the choice of cover story, the semantic proximity of the topic meanings to the critical item and to each other, and the precise organization and length of training and test blocks (**Table 1**). Detailed descriptions of each experimental design, including counterbalancing and randomization, timing procedures, and lexical controls, are available in the Supporting Materials.

The training phase of each study required a set of critical items, competitor topic meanings, and a set of close lexical associates of each topic. From these materials, a set of triplets was created, each of which consisted of a critical word and two different topic meanings. One of these topic words was designated the *preceding* topic, and the other, the *succeeding* topic (**Table 2**).

Triplet	topic**1**	critical	topic**2**
	dream	*fugue*	*music*

Table 2: An example of a training triplet taken from Study 2, in which the critical word *fugue* has been paired with the competing topics *dream* and *music*.

For each topic in a given triplet, corpus data were used to generate a ranked list of its lexical associates. These were used to construct training trigrams, which consisted of the critical item and a pair of its topics' close associates on either side of it (**Tables 3 & 4**). These training trigrams were embedded into larger strings, which subjects were incidentally exposed to in training. The precise number of training trials varied by study.

Condition 1			**Condition 2**		
T**1** associate1	*critical*	T**2** associate1	T**2** associate1	*critical*	T**1** associate1
T**1** associate2	*critical*	T**2** associate2	T**2** associate2	*critical*	T**1** associate2
...	*critical*	...	...	*critical*	...
T**1** associate*N*	*critical*	T**2** associate*N*	T**2** associate*N*	*critical*	T**1** associate*N*

Table 3: Abstract representation of the training trigrams for a given critical word and its two topic meanings.

{DREAM, critical, MUSIC}			{MUSIC, critical, DREAM}		
chasing	*fugue*	*listening*	*listening*	*fugue*	*chasing*
lucid	*fugue*	*classical*	*classical*	*fugue*	*lucid*
worthy	*fugue*	*primal*	*primal*	*fugue*	*worthy*

Table 4: Partial training sets in Study 2 for the critical item *fugue* and its topic synonyms *dream* and *music*. In Condition 1 (right), the ordering of associates is reversed from Condition 2 (left).

Post-training, two tests were administered. In the first, a semantic priming task, a prime word was briefly presented on-screen, and subjects were asked to determine whether the following word was a real word in English. Each critical item was tested in combination with its two competitor topics, alternating its position as a prime or target (**Table 5**).

Subsequently, in a semantic similarity rating task, subjects were asked to rate the similarity of various word pairs on a numerical scale, ranging from "unrelated in meaning" to "identical in meaning". Each critical item was alternately paired with its two topics (**Table 6**).

Semantic Priming	topic**1** → critical	critical → topic**1**	topic**2** → critical	critical → topic**2**
	music → fugue	fugue → music	dream → fugue	fugue → dream

Table 5: Example test trials from the semantic priming task.

Semantic Similarity	topic**1** | critical	topic**2** | critical
	music | fugue	dream | fugue

Table 6: Example test pairs from the semantic rating task.

A key point of difference between studies was the frequency of the critical item: **Study 1** employed pseudo-words, **Study 2**, low frequency items, and **Study 3**, a mix of high (HF) and low frequency (LF) items.

Hypotheses

Priming Semantic priming is a classic paradigm for studying representation in semantic memory (Neely, 1991). A general finding is that a target item will be processed more efficiently when it is preceded by a semantically related prime, with the degree of facilitation depending on the relatedness of the pair. For instance, *bread* will be processed more quickly and accurately when it is preceded by *butter* than when it is preceded by *nurse* (Meyer & Schvaneveldt, 1971). When what is studied is the extent to which recently trained associations can facilitate priming—as is the case here—the priming is classed as ***episodic*** (Hayes & Bissett, 1998; McKoon & Ratcliff, 1979). When those associations are indirectly trained, it can be further classed as ***mediated*** (Lowe & McDonald, 2000).

In our studies, a key consideration is that lexical processing is sensitive to temporal contingencies (Deese, 1965). If subjects learn about both the associative and temporal relations between critical items and their topics, then they should be faster and more accurate on lexical decision trials that are consistent with the sequences observed in training. For example, in training sequences in which $Topic_1 \rightarrow Critical \rightarrow Topic_2$, $Topic_1$ should be a better prime to the critical item than $Topic_2$, and $Topic_2$ should be better primed by the critical item than $Topic_1$.

Similarity Similarity judgments can be affected by the dimensions of alignment that are currently deemed salient to the comparison (Nosofsky, 1986; Tversky, 1977). In the domain of perceptual learning, simulations of convergent and divergent schemas indicate that they develop different feature weights, resulting in correspondingly different representations of the similarity space among exemplars (Ramscar et al., 2010).

If distributional learning is also sensitive to how information is structured in time, then the associative relations the critical item develops with its topics over training should similarly depend on the positioning of their associates. When multiple lexical associates serve to predict a critical item, the information structure will be convergent; when the critical item serves to predict multiple lexical associates, the structure will be divergent (see **Figure 1**).

In the convergent case, competition between the lexical associates present in the preceding context should preferentially weight the shared semantic features they have in common with their topic word. By contrast, in the divergent case, weights will be tuned according to co-occurrence rates, which may not select for the most predictive dimensions. Convergent learning should therefore bring the preceding topic into closer alignment in similarity space with the critical item (Dye & Ramscar, 2009).

Study 1

Subjects were told that scientists had intercepted an alien communication that they had managed to partially translate, but needed further help in order to fully decode. Participants were presented with a series of these cryptic messages, and instructed to learn as much as they could about the alien word in the middle. That critical item was always a nonsense word.

The experiment was designed such that each subject completed ten short experimental sessions, comprising both training and test, one after the other. This meant that participants learned about each critical item in individual blocks, rather than learning about multiple items simultaneously. The design was fully randomized, such that the specific pairing of topic meanings with a given critical item varied by participant. At the end of the experiment, results were aggregated across all sessions.

Participants Eighteen Stanford University undergraduates participated for course credit.

Results In the semantic priming task, lexical decision accuracy was at ceiling, averaging 98%. However, differences in response time were apparent. A paired samples t-test revealed that when the critical item served as a prime to one of its topics, subjects were significantly faster at recognizing succeeding topic words than preceding topic words [$t(17)=2.30$, $p=0.017$], with a mean 37 ms advantage. However, this advantage was mediated by the prime type: when the topic words themselves served as primes to the critical item, no difference was observed between the preceding and succeeding topics [$t(17)=0.25$, $p>0.5$].

After completing the priming task, subjects rated the semantic similarity of each critical item and its competitor topics. A sequential learning account suggests that the preceding topic word should become more similar to the critical item over training. In line with this prediction, subjects rated the preceding topic word as significantly more similar to the critical item than the succeeding topic word [$t(17)=2.27$, $p=0.018$]. Non-parametric analyses of the data, with the Wilcoxon signed-ranks test, yielded the same pattern of results.

Study 2

Subjects were told they were taking part in a study testing their ability to distinguish human from artificial intelligence.

On each trial, they were presented with a trigram sequence (**Table 4**), and asked to judge whether those words had come from a text generated by a human or a computer. In this study, critical items were LF words, whose potential topic meanings were plausible synonyms (e.g., the critical item *abscond* was matched with the topic words *hide* and *flee*). The design was counterbalanced such that the position of each topic word was split evenly across participants. Testing was conducted at the end of the full training session.

Participants 43 undergraduates at Indiana University, Bloomington participated for course credit.

Results The test results of Study 2 replicate the pattern of results in Study 1. In the priming task, lexical decision accuracy averaged 86.4% overall and 81.8% for critical items. A dependent samples t-test revealed that when the critical item served as a prime to one of its topics, subjects were faster [$t(42)=-1.73$, $p=0.046$] and more accurate [$t(42)=2.45$, $p=0.009$] at recognizing topic words that had followed that item, compared to those that had preceded it. A by-items analysis produced a similar pattern for speed [$t(27)=1.53$, $p=0.068$] and accuracy [$t(27)=1.85$, $p=0.038$]. This facilitation pattern was not evident when HF topic words primed LF critical items.

After completing the priming task, subjects rated the semantic similarity of each critical item and its competitor topics. Consistent with Study 1, a dependent samples T-test revealed that preceding topics were rated more similar to critical items, both by subjects [$t(42)=2.99$, $p=.002$] and items [$t(13)=2.83$, $p=0.007$]. Non-parametric analyses, with the Wilcoxon signed-ranks test, confirmed the pattern of results.

Study 3

Subjects were told they were taking part in a study on reaction time. Words were presented one by one, and subjects were instructed to make a keyboard response every time they saw an item that was either a *fruit* or a piece of *furniture*. Training trigrams (**Figure 4**) were pseudo-randomly interspersed throughout this text sequence, with the design counterbalanced such that the position of each topic word was split evenly across participants.

To further assess the extent to which the frequency of the critical item might mediate the predicted effects, both HF and LF critical words were chosen, and each pair of topic meanings was assigned to a pair of unrelated critical items, one in each frequency band (e.g., the critical items *jacket* and *repast* were both assigned the same topic pair). Topic meanings were moderately semantically related to each other, but not to either critical item.

As with Study 2, testing was conducted at the end of the full training session.

Participants 26 undergraduates at Indiana University, Bloomington participated for course credit. Two subjects were dropped from the similarity analyses for selecting the same number for every pair.

Results Study 3 largely replicated the pattern of results in Study 1 and 2. However, in the semantic priming task, the locus of the effect was different: Lexical decision accuracy was at ceiling when HF topic words served as targets (98.7%). However, when topic words served as primes to the critical items, a 2 (training position) by 2 (critical item frequency) repeated measures ANOVA revealed main effects of item frequency for accuracy [$F(1,25)=16.86$, $p<0.001$] and RT [$F(1,25)=29.49$, $p<0.001$], and of training position for accuracy [$F(1,25)=3.82$, $p=.061$]: Subjects were faster and more accurate at recognizing HF targets overall, and more accurate at recognizing critical items that had followed that topic in training, compared to those that had preceded it.

Analysis of the similarity ratings data revealed a main effect of training position [$F(1,22)=5.09$, $p=.034$], a main effect of topic frequency [$F(1,22)=10.07$, $p=.004$], and a marginally significant interaction between training position and critical item frequency [$F(1,22)=3.88$, $p=.062$]. Post hoc analyses (Tukey HSD) indicated that, as predicted, LF critical items became more similar to their topic words over training than did HF items. Further, the effect of the training manipulation was mediated by the frequency of the critical item: The preceding topic word was rated as significantly more similar than the succeeding topic word for LF items ($p<.03$), but not for HF items.

Discussion

Priming Results Speakers appear to be finely attuned to the statistical regularities of their language, allowing them to anticipate upcoming linguistic events based on the current input (Pickering & Garrod, 2007). This notion is supported by our priming results in Studies 1 and 2, which indicate that when the critical items served as primes, subjects were significantly faster to respond to topic words whose associates had occurred *after* the critical items in training. This suggests that episodic priming is sensitive not only to temporal contiguity, but also to directionality.[1]

Interestingly, however, when the prime order was reversed, and the topic words served to prime the critical items, the effect disappeared in two of the three studies. The effect thus appears to be mediated by the frequency relationship between primes and targets.

At first blush, the results of Studies 1 and 2 may seem surprising. In semantic priming, a common finding is that while HF targets are responded to more efficiently overall, it is LF targets that typically show greater facilitation from semantically-related HF primes (Becker, 1979)—not HF targets, as in our studies. However, there are important differences between studies that test semantic memory (pre-existing semantic associations in long term memory), and those that test episodic memory (associations developed over the course of study), like ours.

[1] Our results may seem to invite comparison with those reported in associative priming, where the facilitation provided by forwards and backwards priming is frequently indistinguishable (Koriat, 1981; Thompson-Schill et al., 1998). However, the association norms employed in such studies are distinct from the type of association built through temporal co-occurrence patterns (Jones et al., 2006; Lund, Burgess, & Audet, 1996), and are thus not directly comparable to our findings.

While HF words outperform LF words in semantic tasks, and appear to be more broadly accessible in the lexicon, in episodic paradigms, it is LF words that tend to be better recognized and recalled (Gregg, 1976). This is due, at least in part, to the fact that HF words occur in many more contexts than LF words, making them less associable with any given experimental context (Anderson, 1974; Steyvers & Malmberg, 2003).

The studies presented here examined the extent to which recently trained semantic and temporal associations facilitate priming. As with other episodic tasks, LF words should develop stronger associations to other experimental items than HF words (the similarity analyses in Study 3 attest to this). The key consideration is that these associations are directional: For a given item, its connections *to* other words may be distinct from its connections *from* other words (Nelson & McEvoy, 2000). It follows sensibly then that in Studies 1 and 2, the LF critical items served as effective cue to the HF topic words, even when the reverse does not obtain (Ramscar et al., 2014).

Similarity Results Across three studies, critical items were rated as more similar to their preceding topics than their succeeding topics, a finding predicted by previous modeling simulations of convergent and divergent learning schemas. As with the priming results, the effect of this training manipulation was modulated by the frequency of the critical item (Study 3).

General Discussion

Learning is a temporal phenomena, and it can produce asymmetric outcomes depending on how information is structured in time. Such asymmetries have been previously documented in causal reasoning (Waldmann & Holyoak, 1992) and categorization (Ashby et al., 2002; Ramscar et al., 2010; Yamauchi et al., 2002), and are also attested in sequential learning in non-human animals (Chen et al., 2016). The goal of the present research has been to investigate whether these asymmetric effects might be similarly observable in distributional learning from reading. Across three experiments, our results affirm that they are. An obvious next step is to assess whether models that learn distributed semantic representations of words can replicate these findings (following Jones et al., 2006).

An additional theoretical possibility raised here is that linguistic regularities may play different functional roles depending on whether they participate in convergent or divergent schemas. Suggestive evidence has been offered in artificial language experiments: Whereas stable prefixes and their following nouns are better learned, stable suffixes increase the similarity among those nouns, helping them cohere better as a category (Ramscar, 2013; see also Valian & Coulson, 1988). Biases toward prefixing or suffixing may thus represent a trade-off between ease of processing and learnability, with suffixes facilitating the discovery of grammatical categories among young learners (St. Clair, Monaghan, & Ramscar, 2009), and prefixes serving to reduce uncertainty in online comprehension and production (Dye et al., 2017). This proposal is consistent with the finding that in child-directed speech, new words are preferentially introduced in utterance-final positions (Fernald & Mazzie, 1991), which appears to promote the best learning outcomes (Fernald, Thorpe, & Marchman, 2010; Yu & Smith, 2012). In future research, this framework might be extended to address broader typological questions on the forces at work in language change and evolution.

Acknowledgments

This research was funded by an NSF graduate fellowship to MD. Many thanks to Colin Allen, Brendan Johns, the Shiffrin lab group, and three anonymous reviewers for comments and discussion.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6(4), 451–474.
Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3), 463–498.
Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30(5), 666–677.
Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *ACL*, 238–247.
Becker, C.A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 5(2), 252-259.
Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, 39(3), 510–526.
Chen, J., Jansen, N., and Cate, C. (2016). Zebra finches are able to learn affixation-like patterns. *Animal Cognition*, 19, 65-73.
Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, 114(3), 356–371.
Davis, T., & Love, B. C. (2010). Memory for Category Information Is Idealized Through Contrast With Competing Options. *Psychological Science*, 21(2), 234–242.
Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J.P. Blevins, & H. Bartos (Eds.) *Perspectives on Morphological Organization: Data and Analyses*. Brill: Leiden.
Dye, M. & Ramscar, M. (2009). No representation without taxation: The costs and benefits of learning to conceptualize the environment. *Proceedings of the 31st Meeting of the Cognitive Science Society*, Amsterdam, NE.
Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211.
Firth, J.R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.
Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.
Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. A*cta Psychologica*, 121(1), 65–80.
Jones, M.N., & Mewhort, D.J.K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1), 1–37.
Jones, M.N., Kintsch, W., & Mewhort, D.J.K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory & Language*, 55(4), 534–552.
Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22), 11850–11857.
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. P*sychological Review*, 104(2), 211–240.
Lawrence, D.H. (1949). Acquired distinctiveness of cues: Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, 39, 770-784.
Levering, K. R., & Kurtz, K. J. (2014). Observation versus classification in supervised category learning. *Memory & Cognition*, 43(2), 266–282.
Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15(4), 838–844.
Lowe, W. & McDonald, S. (2000). The direct route: Mediated priming in semantic space. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
McDonald, S. & Ramscar, M. (2001) Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* University of Edinburgh.
Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234.
Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.
Neely, J.H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading: Visual word recognition*, 11, 264-336.
Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory & Cognition*, 28(4), 509–522.
Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances In Neural Information Processing Systems*, 14.
Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. J*ournal of Experimental Psychology: General*, 115(1), 39–61.
Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, 56, 132–143.
Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, 46(4), 377-396.
Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning. *Topics in Cognitive Science*, 6(1), 5-42.
Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science,* 34(6), 909-957.
Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. The American Psychologist, 43(3), 151–160.
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.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605.
St. Clair, M.C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, 33(7), 1317–1329.
Steyvers , M. & Malmberg, K.J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5), 760-766.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5), 352–373.
Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352.
Yamauchi, T. & Markman, A.B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124-148.
Yamauchi, T., Love, B.C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 585-593.