TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 3 (2011) 303–345

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2010.01111.x

Redundancy in Perceptual and Linguistic Experience: Comparing Feature-Based and Distributional Models of Semantic Representation

Brian Riordan, Michael N. Jones

Department of Psychological and Brain Sciences, Indiana University

Received 31 December 2008; received in revised form 17 July 2009; accepted 1 November 2009

Abstract

Since their inception, distributional models of semantics have been criticized as inadequate cognitive theories of human semantic learning and representation. A principal challenge is that the representations derived by distributional models are purely symbolic and are not grounded in perception and action; this challenge has led many to favor feature-based models of semantic representation. We argue that the amount of perceptual and other semantic information that can be learned from purely distributional statistics has been underappreciated. We compare the representations of three feature-based and nine distributional models using a semantic clustering task. Several distributional models demonstrated semantic clustering comparable with clustering-based on feature-based representations. Furthermore, when trained on child-directed speech, the same distributional models perform as well as sensorimotor-based feature representations of children's lexical semantic knowledge. These results suggest that, to a large extent, information relevant for extracting semantic categories is redundantly coded in perceptual and linguistic experience. Detailed analyses of the semantic clusters of the feature-based and distributional models also reveal that the models make use of complementary cues to semantic organization from the two data streams. Rather than conceptualizing feature-based and distributional models as competing theories, we argue that future focus should be on understanding the cognitive mechanisms humans use to integrate the two sources.

Keywords: Semantic modeling; Word learning; Co-occurrence models; Semantic categorization; Latent Semantic Analysis; Child-directed speech

1. Introduction

The last decade has seen remarkable progress in the development of distributional models of semantic representation. Distributional models build semantic representations from the

Correspondence should be sent to Michael Jones, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405. E-mail: jonesmn@indiana.edu

statistical regularities of word co-occurrences in large-scale linguistic corpora. These models are based on the *distributional hypothesis:* The more similar the contexts in which two words appear, the more similar their meanings (Firth, 1957; Harris, 1970). For example, the word *milk* may be observed in the same contexts as *drink* and *cookie.* As a result, it can be inferred that these words are semantically related. Furthermore, it may be induced that *milk* is similar to other words that appear in similar contexts, such as *juice* or *soda* (even if these words do not directly co-occur with *milk*). On the other hand, *milk* will be much less similar to *rain* because both words rarely appear in the same or similar contexts.

Distributional models typically represent word meanings as vectors, where the vector pattern reflects the contextual history of a word across the training corpus. Represented geometrically, words' vectors are points in a high-dimensional semantic space. Words that are more semantically related tend to cluster closer together in the semantic space. Examples of distributional models in cognitive science include Hyperspace Analogue to Language (HAL; Burgess, 1998), Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007), and the Topic model (Griffiths, Steyvers, & Tenenbaum, 2007). Distributional models have provided accounts of a variety of semantic phenomena, including semantic priming (Burgess, 1998; Jones, Kintsch, & Mewhort, 2006; Lowe & McDonald, 2000), word association (Griffiths et al., 2007; Sahlgren, 2006), and semantic categorization (Bullinaria & Levy, 2007; Burgess, 1998; Jankowicz, 2005; Laham, 2000).

Despite their accomplishments, distributional models have been heavily criticized as inadequate psychological models of human semantic learning and representation (e.g., Perfetti, 1998). Critics have argued that meaning cannot be the result of word relationships alone; rather, meaning needs to be grounded in perception and action (for an excellent overview, see de Vega, Graesser, & Glenberg, 2008). For example, proponents of embodied cognition point to a growing body of neuroscientific research demonstrating the activation of sensorimotor areas in the brain in response to words (Barsalou, Santos, Simmons, & Wilson, 2008; Pulvermüller, 2008). Others have demonstrated how distributional models such as LSA fail to represent word and sentence meaning in novel situations because they lack access to the affordances of objects (Glenberg & Robertson, 2000). These affordances can be extracted during language comprehension via perceptual simulations, but they cannot be learned simply from co-occurrence relationships.

These issues have led many to return to classic feature-based theories of semantic representation. In a feature-based model, a word's representation is a binary list of descriptive features (e.g., birds can fly while dogs cannot; see Smith, Shoben, & Rips, 1974). Modern feature-based models of semantic representation are constructed from experiments in which participants generate features that describe words' meanings (McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008). Feature-based models are also susceptible to the criticism of a lack of grounding in real-world referents, as all features are verbal descriptions (Sanford, 2006). However, features can be construed as standing in for sensorimotor experience. For example, part of the meaning of *strawberry* is the perceptual experience of its typical color, which is simply re-encoded in a feature representation as <*is_red*>. Feature-based models have been used successfully to model semantic priming

(Cree, McRae, & McNorgan, 1999; Vigliocco, Vinson, Lewis, & Garrett, 2004) and semantic categorization (Cree & McRae, 2003; Vinson, Vigliocco, Cappa, & Siri, 2003).

Arguing for feature representations' access to most aspects of word meaning, McRae et al. (2005) propose that the feature generation task exploits ''representations that have developed through repeated multisensory exposure to, and interactions with, exemplars of the target category'' (p. 549). Barsalou (2003) hypothesizes that the process of feature generation for a word involves constructing a mental simulation of the object or action, which is then interpreted for verbalizable features.

However, multisensory information from the nonlinguistic environment and linguistic experience are likely to be redundant streams of data, and it is not obvious how language users allocate attention to these two streams. Given an appropriate statistical learner, each source of information may, on its own or in conjunction with the other, support the construction of a psychological semantic space in which semantically similar words cluster together. There is growing evidence that much perceptual information is encoded in language (Connell & Ramscar, 2001; Kintsch, 2007, 2008; Laham, 2000; Louwerse, 2007, 2008; Louwerse, Cai, Hu, Ventura, & Jeuniaux, 2006; Louwerse & Jeuniaux, 2008; Louwerse & Zwaan, 2009), reinforcing the proposal that distributional statistics in language contribute to the representation of word meaning. Assuming that language mirrors the environment it describes, it makes intuitive sense that a great deal of the same information could be learned by attending to either perceptual or distributional input.

The redundancy between distributional and experiential data is core to Louwerse's (2007) *symbol interdependency hypothesis*, an attempt to reconcile the role of distributional statistics in language with the role of embodied simulations in language comprehension. Louwerse allows that many words, functioning as symbols standing for entities in the physical world, must be grounded through embodied semantic representations. However, words as symbolic representations also develop and maintain rich connections with each other. Thus, words may have grounded representations, but do not necessarily. Louwerse proposes that this network of symbolic connections supports efficient online language processing by allowing listeners to skip the enactment of embodied simulations. That is, if distributional structure is available, then language comprehension can rely solely on semantic representations, not needing to invoke perceptual simulations unless deeper semantic processing is required. A similar proposal is made by Barsalou et al. (2008)—cf. Logan's (1988) classic instance theory of learning. Louwerse and colleagues have used LSA to demonstrate that distributional linguistic information encodes a surprising amount of information that is typically thought to require embodied interaction with the extralinguistic environment (Louwerse, 2007, 2008; Louwerse & Jeuniaux, 2008; Louwerse & Zwaan, 2009; Louwerse et al., 2006).

In this article, we extend this line of investigation of the semantic information available in distributional linguistic statistics in four ways. First, we explicitly compare the semantic structure derived by distributional and featural models using a common task. Do featural and distributional models produce similar semantic structure, indicating redundancy in the information available to both types of models? Second, we employ an extensive set of noun and verb concepts. Rather than focusing on whether a specific type of semantic information is encoded in distributional statistics (e.g., geographic information; Louwerse & Zwaan,

2009), we explore whether distributional models are able to derive representations for words across a variety of semantic categories. Third, we examine the representations produced by a variety of distributional models, covering most existing approaches. Do all distributional models encode essentially the same statistical semantic information, or are there nontrivial differences between models? Fourth, we address the acquisition of semantic representations. We investigate the extent to which the distributional statistics of child-directed speech encode semantic information, with the aim of determining whether such statistics might provide a sufficient basis for young learners to acquire and refine word meaning.

In order to compare the different kinds of semantic models, we require a task to which each model can be straightforwardly applied. As noted above, both featural and distributional models have been used to model human data in a variety of semantic tasks. However, the semantic data that the models have attempted to account for is arguably the result of both semantic structure and a process that operates on this structure (e.g., semantic priming, discourse comprehension, etc.). Although both kinds of models can be augmented with processing machinery to model semantic data (e.g., Cree et al., 1999; Johns & Jones, 2009), it then becomes difficult to tease apart the independent contributions of structure and process.

We opt rather to explore the question of how much semantic structure is directly encoded in the representations of featural and distributional models, prior to the addition of a process mechanism. To evaluate this structure, we turn to the task of clustering semantically related words into labeled superordinate semantic classes. For example, *banana* is classified as a Food noun in the WordNet semantic taxonomy (Fellbaum, 1998), while *porcupine* is an Animal noun. Our evaluation focuses on the degree to which the featural and distributional models produce the same assignments of words to semantic classes as made by existing hand-crafted semantic taxonomies.[1] A clustering evaluation allows similarity structure *within* a semantic class to vary while still requiring that differences *between* semantic classes remain observable.

Evaluating the models using clustering also allows analysis of representational content. We evaluate the distinguishing features of particular semantic *clusters* from each model type. Our interest is in exploring the following: (a) the particular features that are most responsible for producing superordinate semantic structure and (b) the extent that such features overlap between the feature norms and distributional models. Overlap in the distinguishing features of similar semantic clusters across feature norms and distributional models may indicate that information relevant for lexical/conceptual representation is redundantly encoded in linguistic and perceptual experience.

An evaluation of featural and distributional model representations should address the question of how these two sources of data—perceptual and linguistic experience—may be used to acquire lexical representations. Children learning their first language may, early on, show a strong reliance on sensorimotor information in acquiring word meanings (Bloom, 2000; Gentner, 2006). However, if much perceptual information is encoded in language, there may be useful statistical clues to meaning and semantic relatedness in children's ambient linguistic environment. For example, the dominant view as to why children learn many nouns before starting to learn verbs is based on perceivability: Nouns have concrete, imageable, and often manipulable referents that remain stable across time (Gillette, Gleitman, Gleitman, & Lederer, 1999). However, Jones and Mewhort (2007)

demonstrated that the statistical structure of language also favors learning nouns before verbs, in the absence of any perceptual information. This is not to say that sensorimotor information is unimportant in acquiring word meanings, but rather that the role of distributional information in the child's linguistic environment may be underappreciated.

In summary, for separate sets of nouns, verbs, and child-relevant nouns and verbs labeled with semantic classes, we compare the performance of feature representations to each of the distributional model representations on the semantic clustering task. The different models are compared according to the quality and accuracy of their clustering solutions with respect to the given sets of semantic classes.

We examine the representations of three different sets of feature representations: McRae et al. (2005), Vinson and Vigliocco (2008), and Howell, Jankowicz, and Becker (2005). Each set has a particular emphasis: McRae et al. (2005) focuses on concrete nouns, Vinson and Vigliocco (2008) contains representations for both nouns and verbs, and Howell et al. (2005) is designed to represent the sensorimotor-based word-meaning knowledge of young children. The distributional models considered here span the variety of different types of models that have proven successful in modeling semantic data.

This article is organized as follows. In Section 2, we provide a description of each of the distributional models considered in this investigation. Section 3 compares the McRae et al. (2005) feature representations for concrete nouns to each of the distributional models trained on a large corpus of English text. Section 4 compares the distributional models to the Vinson and Vigliocco (2008) feature representations for object nouns and action verbs. In Section 5, the distributional models are trained on child-directed speech, and the resulting representations are compared with the Howell et al. (2005) sensorimotor feature representations of both nouns and verbs. Finally, in Section 6, we discuss the implications of our findings for theories of semantic learning and representation.

2. Distributional models

2.1. Overview

Distributional models are comprised broadly of two families based on the type of distributional statistics from which they learn. We refer to these as *context word* models and *context region* models.

Context word models such as HAL (Lund & Burgess, 1996) typically represent similarity by computing a word-by-word co-occurrence matrix. This matrix is populated by moving a window over a text corpus and computing the frequency of co-occurrence of words within the range of the window. In context word models, if two words have counts in the same column of the co-occurrence matrix, then they have occurred with the same word as a neighbor in the moving window. When these words' vectors are compared, the similarity of the words in terms of their co-occurrence with the same neighbors is computed. The aggregation of distributional statistics in this way operationalizes semantic similarity as *substitutability* (Miller & Charles, 1991; see discussion in McDonald, 2000; Lowe, 2001): By calculating

the degree to which two words share the same neighbors, we derive an approximate measure of the extent to which they can be substituted in a variety of contexts.

Context region models, exemplified by LSA (Landauer & Dumais, 1997), compute a word-by-context region matrix, where a context region may be a sentence, paragraph, document, etc. Words are considered semantically similar to the extent that they have appeared in the same contexts. Because the number of contexts is usually very large and the co-occurrence matrix sparse, the dimensionality of the matrix is often reduced using mathematical techniques such as singular value decomposition. The dimensional reduction step allows similar representations for words that do not directly co-occur in the same context (an indirect relationship). Semantic similarity in context region models is operationalized in a manner very different from that of context word models. Landauer and Dumais (1997) interpret context regions as ''episodes,'' and co-occurrences as ''stimuli'' or perceptual ''events'' that become associated with the episodes (p. 216). Semantically similar words, then, are those that are characteristic of discourse on (more or less) coherent topics.

Because distributional models are trained on large text corpora, they have access to vast amounts of statistical information representing many aspects of word meaning. However, a potential obstacle in accessing these statistics is the highly skewed distribution of word frequencies in language. In general, a few words are extremely frequent, while most words are relatively infrequent (cf. Zipf's Law; see discussion in Baayen, 2001). High-frequency words are often function words, which have primarily a syntactic rather than semantic role in language. Thus, co-occurrence statistics may be because of words' frequencies in language—which are in turn correlated with words' syntactic functions—as much as semantic relationships between words (Durda & Buchanan, 2008; Lowe, 2000; Rohde, Gonnerman, & Plaut, 2005, unpublished data). To overcome this problem, many distributional models use a lexical association function (Lowe, 2001; Padó & Lapata, 2007) to transform raw co-occurrences statistics. Lexical association functions are statistical techniques that emphasize co-occurrences that are more frequent than would be predicted by chance, under the assumption that these co-occurrences are the most likely indicators of semantic relatedness (Jurafsky & Martin, 2008). Context word models, in particular, differ primarily in their choice of lexical association function. However, a systematic comparison of models incorporating different lexical association functions and the effect of these transformations on the ability to model human semantic data has not been previously attempted.

In this paper, we explore the representations of members of both the context word and context region model families (Table 1). Within the context word model family, we examine a variety of context word models, each with different lexical association functions. Next, we provide a brief description of each of the models.

2.2. Context word models

2.2.1. HAL

The HAL (Lund & Burgess, 1996) counts the frequency that each context word has preceded or succeeded a word of interest within the moving window. Co-occurrence statistics are computed for each one-word increment of the moving window across a text corpus.

Table 1
List of distributional models compared in the current investigation, along with major parameters (See text for descriptions of each parameter type)

Model	Context Specification	Lexical Association Function	Other Distinctive
Context word			
COALS	Window (ramped)	Correlation	
CS-LL	Window	Log-likelihood score	
CS-LO	Window	Log odds ratio	
HAL	Window (ramped)	Vector length normalization	
HiDEx	Window (ramped)	Word frequency normalization	
Positive PMI	Window	Positive Pointwise Mutual Information	
BEAGLE	Sentence		Word order information
Context region			
LSA	See text	Entropy-based	Dimensionality reduction by SVD
Topics	See text		Dimensions = topics

HAL uses a ''ramped'' window, such that the occurrences of words closer to the target word are weighted more heavily in the co-occurrence matrix. HAL does not employ a lexical association function to transform word co-occurrence statistics.[2] HAL has been used to successfully model several different priming phenomena, including distinguishing semantic and associative priming (Lund & Burgess, 1996) and priming of abstract and emotion words (Burgess & Lund, 1997).

2.2.2. HiDEx

HiDEx (High Dimensional Explorer; Shaoul & Westbury, 2006) is very similar to HAL. Its distinguishing feature is its lexical association function: Each count of the co-occurrences of a target word and a context word is divided by the frequency of the target word in the corpus on which the model is trained. Shaoul and Westbury argue that this method reduces the effect of word frequency in the model's representations.

2.2.3. COALS

COALS (Correlated Occurrence Analogue to Lexical Semantics; Rohde et al., 2005) differs from HAL and HiDEx in two important ways. First, although COALS uses a ramped weighting scheme to weight co-occurrences by their distance from the target word in the window, it dispenses with the left/right context word case distinction, treating all co-occurrence statistics within the window equally. Second, in order to reduce the influence of co-occurrences with words of high-frequency, COALS employs a lexical association function based on Pearson's correlation:

$$\text{weight}(t,c) = \frac{Tw_{t,c} - \sum_j w_{t,j} \sum_i w_{i,c}}{\sqrt{\left(\sum_j w_{t,j}\left(T - \sum_j w_{t,j}\right) \sum_i w_{i,c}\left(T - \sum_i w_{i,c}\right)\right)}} \quad (1)$$

where t is a target word, c is a context word, and

$$T = \sum_i \sum_j w_{i,j} \tag{2}$$

Because negative weights carry little information about semantic relatedness, negative correlations are set to 0. Positive correlations are further transformed by applying the square root function. Rohde et al. demonstrate that COALS provides better accounts of a variety of semantic similarity judgment data than LSA and WordNet-based semantic similarity measures.

2.2.4. CS-LL

Contextual Similarity is one of several context word models developed in McDonald (2000). Like COALS, Contextual Similarity records co-occurrences with context words using a single column for each context word. Unlike COALS, HAL, and HiDEx, no weighting for distance in the context window is used. One version of Contextual Similarity, which we will refer to as CS-LL, uses the log-likelihood score (Dunning, 1993) to factor out chance co-occurrences. If a target word and context word are considered nominal variables, their association can be measured using a contingency table:

	Target	Non-target
Context	q	r
Non-context	s	t

where q is the frequency of the co-occurrence of the target and context word, r is frequency of the co-occurrence of non-target words and the context word, s is the frequency of the co-occurrence of the target and non-context words, and t is the frequency of the co-occurrence of non-target words with context words.

The log-likelihood score can be computed by the following formula (Daille, 1996):

$$\begin{aligned} &q \log q + r \log r + s \log s + t \log t \\ &- (q + r) \log (q + r) - (q + s) \log (q + s) \\ &- (r + t) \log (r + t) - (s + t) \log (s + t) \\ &+ (q + r + s + t) \log (q + r + s + t) \end{aligned} \tag{3}$$

The log-likelihood score has been shown to be much less biased by word frequencies in determining the associations of words in text, and it is commonly used in corpus linguistic research. CS-LL has accounted for semantic similarity judgments and semantic and associative priming (McDonald, 2000; McDonald & Lowe, 1998).

2.2.5. CS-LO

A variant of the CS-LL model proposed by Lowe (Lowe, 2000, 2001; Lowe & McDonald, 2000) replaces the log-likelihood score with a different lexical association function, the log odds ratio. We will refer to this model as the *CS-LO* model. CS-LO uses the log odds ratio to estimate whether the frequency of co-occurrence of a target word with a context word is greater than expected by chance. The odds ratio is computed as follows from a contingency table (see the discussion of the CS-LL model above):

$$\text{odds}(c, t) = \frac{q/r}{s/t} = \frac{qt}{rs} \tag{4}$$

where c is the context word and t is the target word. When the odds ratio is greater than 1, the words are positively associated. The odds ratio is estimated using words' frequencies in the corpus. The CS-LO model has been used to model graded and mediated semantic priming (Lowe & McDonald, 2000).

2.2.6. Positive PMI

Bullinaria and Levy (2007) explore the parameter space of context word models, varying a variety of factors, including the number of context words, type of similarity measure, and lexical association function. Across several tasks to measure both semantic and syntactic representations derived by the models, they found that a lexical association function based on mutual information that uses only positive scores was most successful. Pointwise mutual information (PMI), originally introduced in corpus linguistic research by Church and Hanks (1990), is calculated as:

$$\text{PMI}(c, t) = \log \frac{p(c|t)}{p(c)} = \log \frac{p(c, t)}{p(t)p(c)} \tag{5}$$

where t is the target word, c is the context word, and p represents the probability of an event. Bullinaria and Levy retain only positive scores, setting negative values to zero (hence, positive PMI). A context region model implementation of positive PMI has been shown to outperform other distributional models on semantic similarity judgments when scaled up to very large text corpora (Recchia & Jones, 2009).

2.3. Context region models

2.3.1. LSA

Perhaps the most well-known semantic space model, LSA (Landauer & Dumais, 1997), is a context region model, where the context regions are most commonly paragraphs or sections of text of similar size. Like the majority of context word models, LSA uses a lexical association function. The co-occurrence counts after one pass through the corpus are transformed by a combination of the log-weighted frequency of a word in a context region and the entropy of the distribution over the components of the row vector:

$$\text{weight}(t,c) = \left[\log_2\left(w_{t,c}+1\right)\right] \left[1 + \frac{\sum_j p(t,c) * \log_2 p(t,c)}{\log C}\right] \quad (6)$$

where t is the target word, c is the context word, C is the number of contexts, and

$$p(t,c) = \frac{w(t,c)}{\sum_j w(t,c)} \quad (7)$$

(See Quesada 2007 for details.) This transformation tends to weight vector components by their ''informativeness'' for a target word's meaning: When a word's entropy is low over the contexts of the space, the contexts in which it appears may be expected to carry more information about its meaning.

The key step in LSA is computing the (truncated) singular value decomposition (SVD) on the co-occurrence matrix (Martin & Berry, 2007). Landauer and Dumais contend that the dimensions that result from the application of SVD are ''latent'' semantic dimensions, formed by collapsing similar columns in the matrix. The production of these latent dimensions allows for ''indirect co-occurrences,'' so that words that may not have occurred in the *same* contexts but in *similar* contexts will come to have more similar vector representations, and will thus be closer in the semantic space. This, Landauer and Dumais argue, allows LSA to more accurately reflect human-like word similarity judgments, among a range of other semantic memory phenomena (Landauer & Dumais, 1997).

2.3.2. Topic model

The Topic model (Griffiths et al., 2007; Steyvers & Griffiths, 2007) is based on the idea that context regions are collections of words that have been sampled from a mixture of latent topics. Each context region in a corpus is associated with a probability distribution over a set of topics of a chosen size, and each topic specifies a probability distribution over words. The model learns these probabilities by taking account of the statistics of the co-occurrence of words within context regions in the training corpus. Words are similar to the extent that they share the same topics.

Although the Topic model is based on word co-occurrences within context regions, it represents a significant departure from the other distributional models considered in this paper. First, the Topic model is a probabilistic generative model: The model specifies a stochastic process over latent variables by which words in a corpus can be generated. Second, because of the structured representation provided by the conditional dependencies of the model's latent variables, the model is able to represent different senses of words. Griffiths et al. (2007) demonstrate that the Topic model outperforms LSA in predicting word association and several other semantic phenomena.

2.4. Models incorporating word order

Recently, several models have been developed that attempt to enrich the semantic representation derived from a word's distributional profile in linguistic experience with

information about the sequential dependencies that tend to hold between a word and other words with which it co-occurs. The result in these models tends to be a representation that blends syntagmatic and paradigmatic information. For example, Griffiths, Steyvers, Blei, and Tenenbaum (2005) have merged the Topic model with a Markov chain process to learn short-range syntactic dependencies. Similarly, Dennis (2005) has integrated paradigmatic and syntagmatic information sources in a multiple-trace memory model based on string edit theory. Rao and Howard (2008) have also successfully trained a version of the Temporal Context Model to infer semantic relationships by predicting a word's temporal context. The model we focus on here is the recent BEAGLE model of Jones and Mewhort (2007).

2.4.1. BEAGLE

BEAGLE is based on an alternative approach to constructing distributional representations called *random indexing* (cf. Kanerva, Kristoferson, & Holst, 2000; Sahlgren, 2006). This method begins by assigning random vectors to words. As each word in the corpus is processed, the random vectors of other words in a given context are added to the word's memory vector. At the end of learning, a word's memory vector is a pattern of elements that represents the sum of its distributional experience with other words.

BEAGLE is similar to context word models in that co-occurrence statistics are computed relative to other words. BEAGLE combines this random indexing algorithm with a method for encoding the history of the order in which a word appears with respect to other words. Thus, BEAGLE attempts to capture both semantic information (modeled by co-occurrences) and linear order information in the same distributed representation (see also Sahlgren, Holst, & Kanerva, 2008). Order information is modeled in BEAGLE by keeping track of the *n*-grams of various sizes in which a word appears; *n*-gram information is compressed into a single representation using vector convolution. BEAGLE has been used to model a number of phenomena, including semantic priming, typicality, and semantic constraints in sentence completions (Jones & Mewhort, 2007; Jones et al., 2006).

3. Comparing feature-based and distributional representations of concrete nouns

Much research on semantic memory focuses on the representation of concrete nouns. Our first analysis compares the representations of the distributional models for this class of words with the McRae et al. (2005) feature norms, which provide representations for a variety of concrete nouns.

3.1. The McRae et al. (2005) feature norms

The McRae et al. (2005) feature norms (hereafter MCSM) were designed to provide feature representations for the meanings of words commonly used in studies of semantic memory and semantic impairments. The norms include participant-generated feature representations for 541 concrete nouns. Participants were instructed to come up with a variety of features, including those referring to perceptual, functional, encyclopedic, and

taxonomic information. Features were produced by 30 participants for each word (McRae et al., 2005). Only features produced by at least 5 of 30 participants were retained, resulting in a total of 2,526 features describing all concepts. Table 2 provides an example of the features in the MCSM representations for the concept *bear.* The MCSM feature representations have been used to account for a variety of human behavior related to semantic representation, including semantic priming (Cree et al., 1999; McRae, de Sa, & Seidenberg, 1997), feature verification (McRae, Cree, Westmacott, & de Sa, 1999), and trends in semantic category deficits (Cree & McRae, 2003).

3.2. Method

3.2.1. Semantic classes

Because the MCSM feature norms do not provide semantic classes for each word, we used semantic class information derived from WordNet. The classes we used are the *lexicographer files (lex-files).* These files are coarse-grained semantic labels used by lexicographers for organizing the addition of new words to the WordNet hierarchy (Miller, 1998). The lex-files are comprised of 44 broad semantic categories spanning the syntactic classes contained in WordNet: nouns (25), verbs (15), adjectives (3), and adverbs (1). Table 3 lists the classes and the number and kind of words that formed each class.

3.2.2. Distributional model implementation and training

The distributional models were trained on the Touchstone Applied Science Associates (TASA) corpus. This corpus is intended as a ''quantitative summary of the printed vocabulary encountered by students in American schools'' (Zeno, Ivens, Millard, & Duvvuri, 1995). It contains short samples from textbooks, works of literature, and popular works of fiction and nonfiction (Dennis, 2007). The TASA corpus has been commonly used to train distributional models in cognitive science research (Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997).

The TASA corpus was lemmatized and part-of-speech tagged using TreeTagger (Schmid, 1994). Lemmatization involves replacing inflectional variants with a common lexical form. For example, *walked* is replaced with *walk,* and *tigers* is replaced with *tiger.* Following McDonald (2000), we assume that semantically relevant distributional information is preserved across inflectional variants. Thus, for example, the distributions of *tiger* and *tigers* are assumed to provide similar information for the distributional learners. Lemmatization is widely used in information retrieval and computational linguistics as a preprocessing step in order to improve the statistical foundation of models: Collapsing across inflectional variants increases the frequency of word lemmas, thereby allowing more distributional information to be incorporated into model representations. However, we do not assume the same meaning is maintained across syntactic categories. For example, we distinguish $break_{noun}$ from $break_{verb}$. For this reason part-of-speech tagging was employed. After preprocessing, TASA contained 10,625,017 tokens.

The implementation of the context word models involved some simplification of the models as reported in the literature. First, a uniform, automatic procedure was employed to

Table 2
Example features for the word *bear* from McRae et al. (2005) (MCSM), Vinson and Vigliocco (2008) (VV), and Howell et al. (2005) (HJB)

MCSM		VV		HJB	
is_large	23	big	14	breathes	0.98
has_fur	20	fur	14	drinks	0.98
an_animal	19	animal	13	eats	0.98
has_claws	15	brown	9	has_4_legs	0.98
is_black	14	black	7	has_eyes	0.95
is_brown	13	wood	7	has_face	0.95
beh_-_eats_fish	9	claw	6	has_legs	0.95
is_white	9	mammal	6	has_paws	0.95
beh_-_hibernates	8	scare	5	has_teeth	0.95
has_feet	8	4-legs	4	has_fur	0.93
hunted_by_people	8	danger	4	has_tail	0.90
has_teeth	7	wild	4	scariness	0.88
lives_in_wilderness	7	ferocious	3	crawls	0.85
beh_-_produces_offspring	6	grow	3	makes_animal_noise	0.85
lives_in_forests	6	paw	3	size	0.85
has_4_legs	5	teeth	3	strength	0.85
has_paws	5	walk	3	weight	0.85
is_dangerous	5			runs	0.83
				climbs	0.80
				width	0.80
				swims	0.78
				noise	0.73
				temperature	0.73
				height	0.70
				is_black	0.70
				is_solid	0.68
				length	0.68
				speed	0.68
				beauty	0.63
				roughness	0.63
				intelligence	0.60
				is_brown	0.60
				is_strong_smelling	0.60
				colourfulness	0.55
				has_whiskers	0.55
				is_old	0.55
				goodness	0.50
				is_lovable	0.50

Note. In the Vinson and Vigliocco example, only features for which the production frequency was 3 or greater for this word are shown. In the Howell et al. example, only features which received an average rating of greater than 0.5 for this word are shown.

choose the context words. The procedure approximated a specified number of *content* words (nouns, verbs, and adjectives), considering words by their frequency in the corpus in descending order. A short stop word list was used to exclude the verbs *be, have, do,* and

Table 3
WordNet semantic classes for concrete nouns from McRae et al. (2005)

Semantic Class	Number of Words	Example Words
Animals	136	Eagle, panther, spider, clam
Artifacts	325	Coat, jar, ambulance, pistol
Foods	66	Blueberry, spinach, cheese, shrimp
Objects	5	Beehive, emerald, pearl, rock
Plants	8	Pine, dandelion, vine, stick

quantifiers such as *many, few,* and a few other semantically light high-frequency adjectives (*other, such*). The window size of the context word models was fixed at 10 words, and its shape was symmetric, centered on a word of interest. That is, the models computed co-occurrences over a window of five words to both the left and right of a target word. BEAGLE was trained on the same corpus, using sentences as contexts in which to accumulate co-occurrence and order information.

Latent Semantic Analysis and the Topic model were trained on a version of the same lemmatized and part-of-speech-tagged TASA corpus that maintained the context regions from the original corpus. In this and the following analyses, we experimented with several reduced dimensionalities for LSA in order to find the best LSA model to compare with the other distributional models. For the LSA model representations, the best-performing dimensionality on the clustering task was 100.[3]

Parameters for the Topic model were similar to those reported in Griffiths et al. (2007).[4] Markov-chain Monte Carlo (MCMC) sampling was carried out with three Markov chains run for 1,600 iterations each, with $\alpha = 50/T$ (T = the number of topics) and $\beta = 0.01$. After discarding the first 800 iterations as burn-in, each chain was sampled once every 100 iterations, for a total of eight samples per chain. Simulations were run with 300 topics.[5] Vector representations for each word were calculated from the probability distribution of each word over the topics (Griffiths et al., 2007).

3.2.3. Target words

Of the 541 words in the MCSM norms, 470 were used in the analysis. We removed homonyms, words unknown to the part-of-speech tagger, and words not contained in WordNet. A further criterion for inclusion was a frequency of 10 or greater in the TASA corpus; 41 words did not meet this criterion and were excluded.

3.2.4. Clustering

The MCSM featural representations and the distributional models were evaluated on how well semantically similar concrete nouns from the McRae et al. (2005) feature data clustered together in each model's high-dimensional semantic space. We measured how well each type of model was able to cluster words with common semantic class labels from WordNet into semantically coherent clusters using the similarity encoded in each model's representations.

In the clustering paradigm employed here, the experimenter requests a fixed number of clusters. The clustering algorithm then uses the data to produce a solution with the requested number of clusters that maximizes within-cluster similarities and between-cluster differences. An ideal solution in the current case, then, would be five clusters, where each cluster contains words from only one semantic class. Note that the number of clusters requested is the number of different semantic category labels in WordNet that cover the words in the dataset.[6]

Clustering solutions were derived for the MCSM representations and representations from the nine distributional models. Clustering was performed using CLUTO (Karypis, 2003).[7] The clustering algorithm we used was partitional: Starting with a single cluster, the data are repeatedly bisected into more specific clusters, and the clusters are refined by means of a global criterion function. The criterion function used was I2, which maximizes the similarity of vectors within a cluster. Partitional clustering with the I2 criterion function was chosen because this combination has been shown to consistently outperform agglomerative clustering methods and other partitional methods (Zhao & Karypis, 2002). We used the cosine similarity function.

Because CLUTO uses a randomized greedy incremental cluster refinement algorithm, clustering solutions usually vary in their goodness (Zhao & Karypis, 2001). Therefore, for each model except the Topic model, we ran the clustering algorithm 20 times and report means and standard errors for summary statistics on each model's set of solutions.[8] For the Topic model, summary statistics were computed once from each sample of each chain, for a total of 24 solutions. The summary statistics computed for clustering solutions were *purity and entropy* (Zhao & Karypis, 2001).

The *purity* of a cluster measures the fraction of the cluster size that the largest semantic class within the cluster represents. If a cluster contains only one semantic class, then its purity is 1. Cluster purity for a cluster c_r of size n_r is calculated as:

$$\text{Purity}_{\text{cluster}}(c_r) = \frac{1}{n_r}\max_i\left(n_r^i\right) \tag{8}$$

The purity of a clustering solution is the sum of the purities of the clusters, weighted by cluster size:

$$\text{Purity}_{\text{solution}} = \sum_{r=1}^{k}\frac{n_r}{n}\,\text{Purity}_{\text{cluster}}(c_r) \tag{9}$$

Larger values of purity indicate better clustering solutions.

The *entropy* of a cluster measures the extent to which semantic classes vary within a cluster. If there is no variation of semantic classes within a cluster, then entropy is 0.

Cluster entropy for a cluster c_r of size n_r is calculated as:

$$\text{Entropy}_{\text{cluster}}(c_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{10}$$

where q is the number of classes in the data and n_r^i is the number of words in the ith class that were assigned to the rth cluster. The entropy of a clustering solution is the sum of the entropies of the clusters, weighted by each cluster's size:

$$\text{Entropy}_{\text{solution}} = \sum_{r=1}^{k} \frac{n_r}{n} \text{Entropy}_{\text{cluster}}(c_r) \tag{11}$$

where there are k clusters and n total words. *Smaller* values of entropy indicate better clustering solutions.

3.3. Results and discussion

3.3.1. Model comparison

Fig. 1 presents the clustering solution purity and entropy scores of each model. Pairwise t tests with Bonferroni corrections on the purity scores revealed that the MCSM features ($M = 0.879$, $SD = 0.003$) were significantly *lower* than each of the

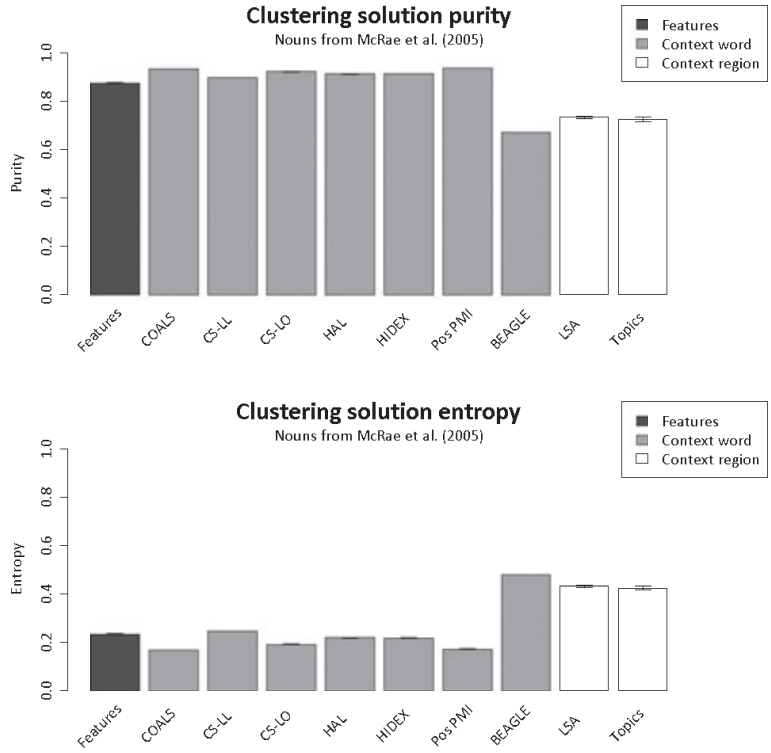

Fig. 1. Clustering solution purity (top) and entropy (bottom) scores for the featural and distributional models on McRae et al. (2005) noun dataset. Error bars indicate standard errors. Good clustering solutions are characterized by high purity and low entropy. Colors indicate the different classes of models: the feature-based model, context-word distributional models, and context-region distributional models.

context word distributional models except BEAGLE (all $p < .0001$). That is, each of these context word distributional models produced clusters that contained a greater number of words from a single semantic class than did the MCSM feature representations. The models with the highest purity scores were COALS ($M = 0.936$, $SD = 0.000$), CS-LO ($M = 0.924$, $SD = 0.003$), and positive PMI ($M = 0.939$, $SD = 0.002$), which did not differ significantly from each other. Although BEAGLE is a hybrid model, it tended to perform much like context region models on the clustering task, and worse than the context word models.

The results for the entropy scores were similar. Each of the context word distributional models' entropy scores except for CS-LL was significantly lower than the MCSM features ($M = 0.237$, $SD = 0.004$), all $p < .0001$. COALS ($M = 0.170$, $SD = 0.001$) and positive PMI ($M = 0.173$, $SD = 0.006$) had the lowest entropy scores and did not differ significantly.

These results demonstrate that the distributional statistics in the linguistic environment are sufficient to derive robust representations of semantic similarity. The representations derived by a subset of models, specifically context word models, appear comparable to (if not better than) the structure available in human-generated feature representations.

The performance of many of the distributional models rivaled that of the MCSM feature representations, but does this mean that the two types of models used similar information to arrive at their respective clustering solutions? To address this question, we undertook a detailed analysis of clustering solutions from the MCSM representations and one of the best-performing distributional models, COALS.

3.3.2. Comparison of MCSM and COALS representations

We selected each model's best clustering solution in terms of purity and entropy for further analysis (analogous to choosing the maximum likelihood fit of the models to the data). The MCSM solution had a purity of 0.879 and an entropy of 0.235, while the COALS solution had a purity of 0.926 and an entropy of 0.170.[9] First, each cluster in each solution was labeled with the semantic class of the majority of words in the cluster. These labels, along with the purity and entropy of individual clusters,

Table 4
Comparison of the clustering solutions of COALS and the MCSM feature representations by semantic class for the nouns from McRae et al. (2005) (Note that cluster numbers begin with 0)

COALS				MCSM			
Cluster	Semantic Class	Entropy	Purity	Cluster	Semantic Class	Entropy	Purity
3	Animals	0.220	0.922	0	Animals	0.173	0.936
				1	Animals	0.259	0.853
0	Artifacts	0.170	0.938	3	Artifacts	0.068	0.977
2	Artifacts	0.132	0.945	4	Artifacts	0.097	0.970
4	Artifacts	0.088	0.975				
1	Foods	0.330	0.847	2	Foods	0.692	0.593

are shown in Table 4. Both the MCSM representations and COALS find clusters with majority labels covering three of the five semantic classes.

Both the MCSM and COALS solutions include reasonable clusters for the semantic classes of Animals and Artifacts. MCSM splits the Animals class into two clusters: flying animals, including both birds and insects (cluster 0), and all other animals (cluster 1). As Table 5 shows, the important features for the flying animals cluster are <*a_bird*> and <*beh_-_flies*>,[10] while the <*an_animal*> feature describes the other animals cluster. COALS, on the other hand, includes all animals in one cluster (3), which is characterized by co-occurrence with words such as *eat, animal,* and *fly.* Both MCSM and COALS make use of similar features—*animal, fly*—suggesting these are encoded redundantly in perceptual and linguistic experience.

The Artifacts class is also split into several clusters by both models. MCSM divides artifacts into two clusters based on what they are made of (Table 5): metal (cluster 3) and wood (cluster 4). COALS separates the artifacts into three clusters (0, 2, 4). As the descriptive features of these clusters in Table 5 show, these clusters represent some of the things that humans do with objects: wearing (cluster 0: articles of clothing), riding (cluster 2: vehicles), and holding (cluster 4: manipulable objects). Thus, MCSM and COALS use very different criteria for grouping artifacts. According to MCSM, <*made_of*> properties are the most salient features for human semantic similarity. COALS, on the other hand, organizes artifacts by the type of actions that they undergo or are used for.

Both models also produce a single cluster for the Foods semantic class. For the MCSM vectors, the feature <*is_edible*> binds together the words that form the Foods cluster (cluster 2). COALS' Foods cluster is organized around things humans do with food (*eat, grow*) and superordinate categories associated with them (*fruit*). Hence, COALS is able to find the most important feature used by the MCSM vectors, <*is_edible*>, in the co-occurrence statistics of the corresponding verb, *eat*.

Are there differences in the global structure of the semantic spaces formed by the models? The clustering solutions for MCSM and COALS are presented visually in Fig. 2. Here, multidimensional scaling has been performed on the solutions to represent the similarities between clusters in two dimensions. Each cluster is represented as a peak. The height of each peak represents the cluster's internal similarity, while the area of the peak represents the number of elements in the cluster (Rasmussen, Newman, & Karypis, 2003). Above each cluster are the three features from the data that are most descriptive of the cluster.

In the MCSM solution (top panel), we see the two animal clusters (0, 1) in close proximity. The two artifacts clusters (3, 4) are shown to be quite heterogeneous, with minimal rise in the surface of the cluster landscape at their center points.

In the COALS visualization (bottom panel), the three artifact clusters appear in the middle and to the right (0, 2, 4). Unlike in the MCSM multidimensional scaling solution, these clusters have relatively robust internal similarity. The clothing cluster (0) is particularly homogeneous. Thus, according to the COALS MDS solution, the global semantic space of concrete nouns is well-described by the three semantic classes and five clusters from its clustering solution.

Table 5
Comparison of the most descriptive and discriminating features for the clustering solutions of COALS and the MCSM feature representations on the nouns from McRae et al. (2005) (Note that cluster numbers begin with 0)

COALS						MCSM					
Cluster	Semantic Class	Descriptive Feature	% Within-Cluster Similarity Explained	Discriminating Feature	% Between-Cluster Similarity Explained	Cluster	Semantic Class	Descriptive Feature	% Within Cluster Similarity Explained	Discriminating Feature	% Between-Cluster Similarity Explained
3	Animals	say-V	0.04	animal-N	0.04	0	Animals	a_bird	0.33	a_bird	0.20
		eat-V	0.03	fly-V	0.03			beh_-_flies	0.21	beh_-_flies	0.13
						1	Animals	an_animal	0.46	an_animal	0.25
								is_large	0.08	made_of_metal	0.10
0	Artifacts	wear-V	0.15	wear-V	0.22	3	Artifacts	made_of_metal	0.49	made_of_metal	0.27
		blue-J	0.05	blue-J	0.05			has_a_handle	0.06	an_animal	0.07
2	Artifacts	ride-V	0.03	ride-V	0.05	4	Artifacts	made_of_wood	0.32	made_of_wood	0.18
		go-V	0.02	eat-V	0.03			is_long	0.08	an_animal	0.08
4	Artifacts	put-V	0.02	eat-V	0.03						
		hold-V	0.02	hold-V	0.03						
1	Foods	fruit-N	0.07	fruit-N	0.10	2	Foods	is_edible	0.22	is_edible	0.12
		eat-V	0.06	grow-V	0.07			is_green	0.15	made_of_metal	0.10

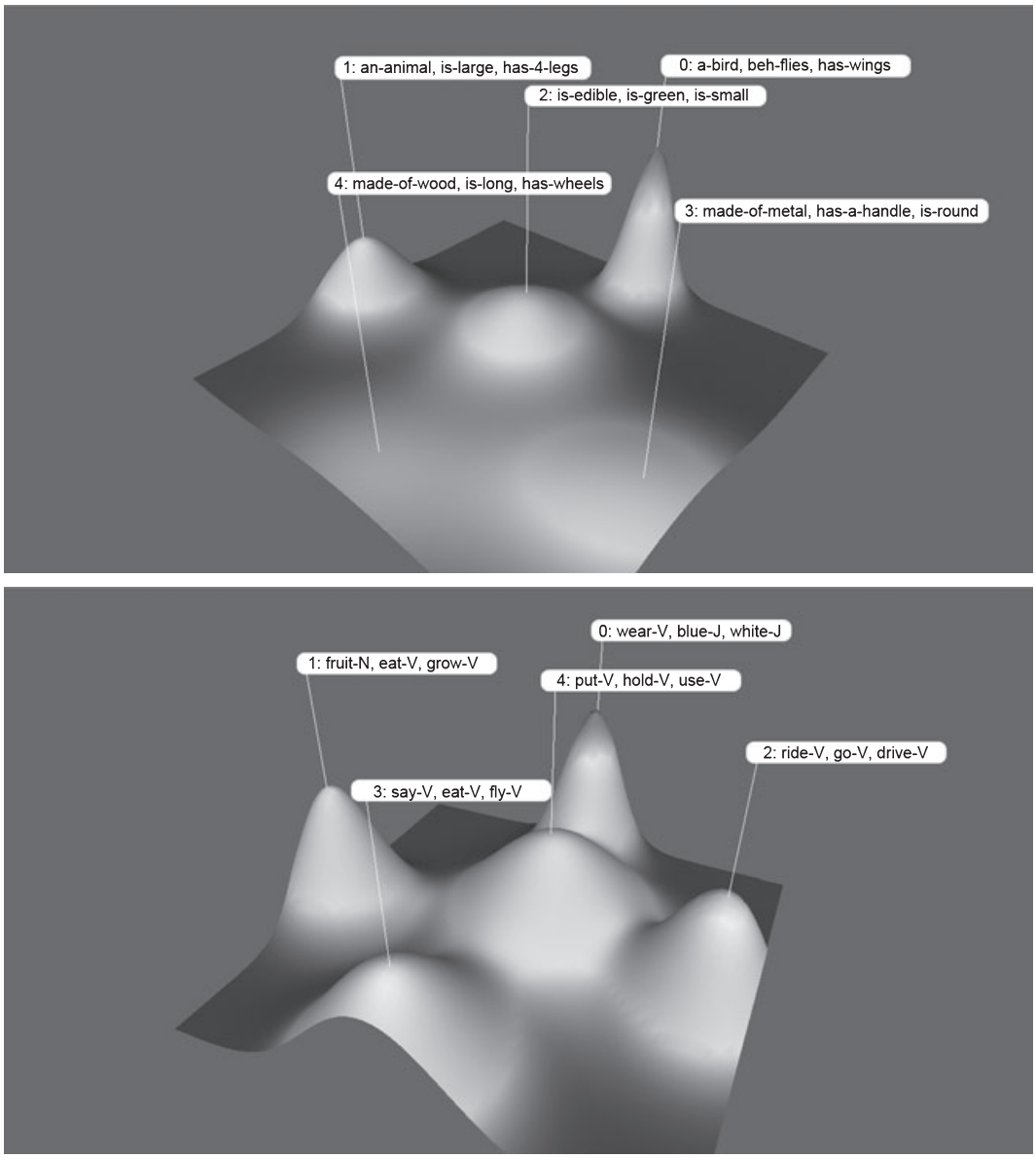

Fig. 2. Multidimensional scaling solution for an example 5-cluster solution from MCSM (top panel) and COALS (bottom panel) for nouns from McRae et al. (2005). Peaks are labeled with cluster numbers and the three most descriptive features for each cluster.

To summarize, the semantic similarity structure of the MCSM representations is governed by different types of features depending on the semantic class. For example, animals are characterized by behaviors (<*flies*>) and taxonomic features (<*is_animal*>),

while artifacts are best described by <*made_of*> properties. COALS, on the other hand, derives semantic similarity for concrete nouns based on the co-occurrence of shared, frequent action verbs *(eat, fly).* These verbs tend to be functions and behaviors of the nouns.

While there is no reason in principle that a distributional model such as COALS could not find a perceptual feature such as <*made_of*>, this information is probably much less frequent in the distributional statistics of linguistic experience, and so less likely to form an important part of distributional models' representations. Nevertheless, the distributional information that is present allows for the creation of very plausible semantic clusters. In the case of the concrete nouns used by McRae et al. (2005), we have seen that the representations of several distributional models are comparable in their clustering ability to that of the MCSM vectors. Furthermore, detailed analysis of a COALS clustering solution demonstrated that the clusters it produced were qualitatively plausible as well. In addition, COALS predicts that the organization of artifacts will tend to be governed more by action (i.e., their functions) than by perceptual similarity.

4. Comparing feature-based and distributional representations of object nouns and action verbs

The previous section demonstrated that distributional models trained on a reasonably large and representative corpus can derive semantic similarity structure that is as robust as that extracted from feature representations produced by human participants. One question that arises is whether this result is because of the particular information included in the McRae et al. (2005) data, for example, the participants in the norms and the particular methods used to collect the data. Another question is related to the limited scope of the MCSM representations: Can distributional models learn plausible semantic similarity relationships for other kinds of words such as verbs?

The analysis presented in this section sought to address both of these questions. We compared the distributional models' representations to the feature representations of Vinson and Vigliocco (2008), which include both concrete objects and event nouns as well as action verbs.

4.1. The Vinson and Vigliocco (2008) feature norms

The Vinson and Vigliocco (2008) feature norms (hereafter VV) contain representations for 456 words of three types: objects (169), action verbs (216), and event nouns (71). Action verbs include both concrete (*hit, sneeze*) and more abstract (*lose, suggest*) concepts. Event nouns, too, include both concrete (*snore*) and abstract (*challenge*) events. As with the MCSM feature norms, features were generated by university students in a laboratory setting. Participants were asked to describe a selection of words using features. Features were produced by 20 participants for each word. Features produced less than nine times across all words were discarded, resulting in 1,029 features

Table 6
Example features for the word *eat* from Vinson and Vigliocco (2008) (VV) and Howell et al. (2005) (HJB)

VV		HJB	
food	17	mouth	1.00
consume	9	perception_taste	0.98
mouth	8	tongue	0.98
hunger	7	decreases_hunger	0.97
nutrition	7	lips	0.93
swallow	7	increases_energy	0.87
daily	3	purposeful	0.85
digest	3	fingers	0.80
humans	3	pleasureable	0.75
ingest	3	wrist	0.75
survive	3	consumes	0.72
		face	0.72
		perception_touch	0.72
		head	0.67
		perception_smell	0.67
		amount_of_contact_involved_between_actor_and_object	0.65
		requires_physical_object	0.65
		increases_thirst	0.62
		balance	0.60
		requires_a_specific_overall_bodily_position	0.60
		attention	0.58
		involves_container_containing	0.58
		perception_visual	0.58
		awareness	0.57
		reactiveness	0.57
		neck	0.55
		control	0.52
		degree_of_overall_body_contact_involved	0.52
		ends_something_else	0.52
		eyes	0.50
		transference_of_something	0.50

In the Vinson and Vigliocco example, only features for which the production frequency was 3 or greater for this word are shown. In the Howell et al. example, only features which received an average rating of greater than 0.5 for this word are shown.

across all words. Thus, a word's representation is a vector of length 1,029 containing the frequency of production of each feature for a given word. An example of the VV feature representation for the word *bear* is given in Table 2. Table 6 shows the VV feature representation for the verb *eat.* The VV feature database has been used to make predictions regarding types of semantic impairments (Vinson et al., 2003), performance in psycholinguistic tasks such as picture-naming (Vigliocco, Vinson, Damian, & Levelt, 2002; Vigliocco et al., 2004), and the role of syntactic factors in semantic representation (Vigliocco et al., 2005).

Table 7
WordNet semantic classes for object nouns from Vinson and Vigliocco (2008)

Semantic Class	Number of Words	Example Words
Animals	28	Donkey, fish, paw, tail
Artifacts	75	Blouse, pen, helicopter, gun
Body	23	Arm, mouth, tooth, hair
Foods	31	Pineapple, lemon, cabbage, bean

Table 8
VV semantic classes for action verbs from Vinson and Vigliocco (2008)

Semantic Class	Number of Words	Example Words
Body actions	36	Blink, grin, eat, kick
Body sense	11	Smell, taste, feel, ache
Change of location	8	Move, pull, put, send
Change of state	10	Fill, mix, pour, twist
Communication	25	Demand, read, teach, shout
Construction	5	Build, construct, fix, paint
Contact	6	Bump, crash, hit, press
Cooking	7	Bake, boil, fry, steam
Destruction	8	Break, kill, chop, smash
Exchange	15	Accept, borrow, pay, want
Heat/light emission	6	Flame, flicker, shine, sparkle
Motion direction	10	Arrive, descend, escape, rise
Motion manner	23	Walk, dive, fly, bounce
Noises	8	Clang, rattle, sigh, sing
Noise-animal	4	Bark, chirp, meow, growl
Tool action	8	Brush, hammer, pound, shovel

4.2. Method

4.2.1. Semantic classes

The Vinson and Vigliocco (2008) feature database includes semantic class information for both nouns and verbs. For the noun dataset, however, we used WordNet semantic classes in order to make the clustering results more comparable to those for the MCSM nouns. Table 7 provides a list of the semantic classes for the nouns. In the VV dataset, there are four classes, including a class for parts of the body.

We used the VV semantic classes in the verb cluster evaluation. Table 8 lists the semantic classes and gives the distribution of verbs across these classes. A total of 16 classes are used to classify the verbs in the dataset. The classification is relatively fine-grained: There are multiple classes for Body, Motion, and Noises.[11]

4.2.2. Distributional model training and implementation

The implementation and training of the context word models was identical to the previous analysis. For the LSA model representations, the best-performing dimensionality was 100 for the nouns and 300 for the verbs. The Topic model was run with 300 topics.

4.2.3. Target words

In order to compare the results for the VV feature representations with the results for the McRae et al. analysis, we did not include the event nouns in this analysis. Because feature model representations tend to be focused on a single word sense (see McRae et al., 2005 for discussion), we sought to make this the case for distributional model representations. However, because the TASA corpus is not tagged by word sense, distributional models conflate the distributions of all word senses for a given lemma. If many different word senses for a word are present in the corpus, the co-occurrence information in word's vector representations will aggregate the information for all senses, making them difficult to compare with the feature representations. Therefore, we elected to remove words from the object noun and action verb target set that showed a high degree of polysemy, which we operationalized as 15 or more senses in WordNet. This affected five nouns (e.g., *file, mouse*) and 17 verbs. The resulting target set included 157 nouns and 190 verbs. Of the nouns, 114 were also contained in the McRae et al. data.

4.2.4. Clustering

The same parameters were used to derive clustering solutions as in the previous analysis. For the noun dataset, we derived 4-cluster solutions, matching the number of semantic classes covering the words in the data. For the verb dataset, 16-cluster solutions were requested.

4.3. Results and discussion

4.3.1. Nouns

The purity and entropy scores for each model are shown in Fig. 3. Pairwise t tests with Bonferroni corrections on the purity scores showed that COALS ($M = 0.832$, $SD = 0.012$) and the VV features ($M = 0.843$, $SD = 0.005$) did not differ significantly. Both the feature representations and COALS had significantly higher purity scores than each of the other distributional models (all $p < .0001$). The entropy scores of the VV representations' solutions and those of COALS also did not differ ($p = .147$). The VV feature and COALS model entropy scores were significantly lower than any of the other models (all $p < .0001$).

Two aspects of these results are of note. First, these results serve to replicate the finding from Section 2 that at least some distributional models can produce semantic similarity structure of similar quality to human-generated feature representations. Second, there was a clear difference between the performance of COALS and each of the other distributional models. Because there was significant variability within the context word models, this result must be attributed to the different effects of the context word models' lexical association functions. In this case, the COALS correlation-based method of reducing the impact of chance co-occurrences proved better able to reproduce the WordNet semantic classes.

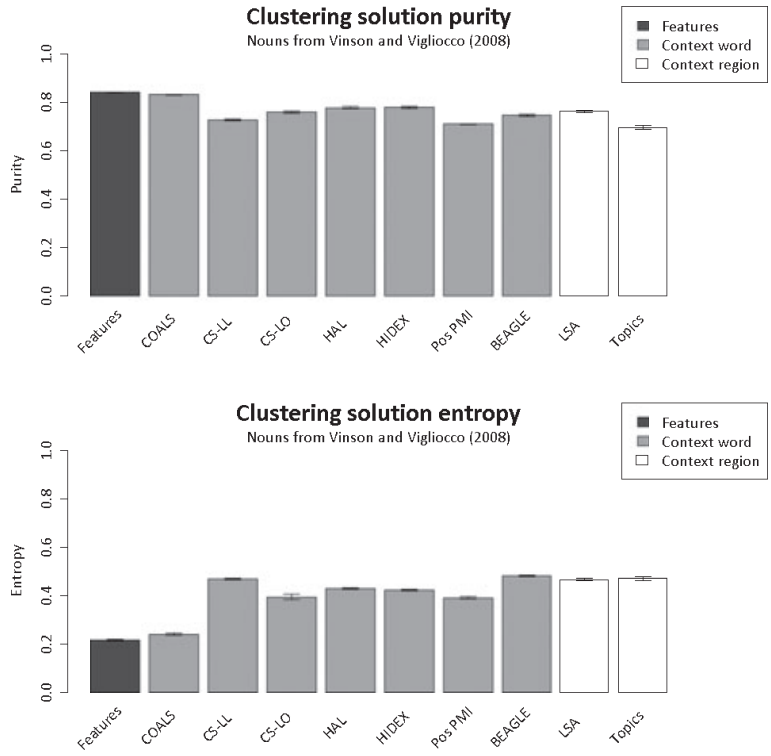

Fig. 3. Clustering solution purity (top) and entropy (bottom) scores for the object nouns from Vinson and Vigliocco (2008).

4.3.2. Verbs

Fig. 4 shows the purity and entropy scores for the models on the Vinson and Vigliocco (2008) verb dataset. The pairwise *t* tests with Bonferroni corrections on the purity scores revealed that the VV features' scores ($M = 0.534$, $SD = 0.012$) were significantly higher than all of the distributional models (all $p < .0001$). Among the distributional models, the models with the highest purity scores were COALS ($M = 0.479$, $SD = 0.029$), HAL ($M = 0.492$, $SD = 0.008$), and HiDEx ($M = 0.489$, $SD = 0.009$), which were not significantly different from each other. With respect to the entropy scores, the VV feature representation solutions' scores ($M = 0.408$, $SD = 0.014$) were the lowest, $p < .0001$. COALS' entropy scores ($M = 0.445$, $SD = 0.014$) were the lowest among the distributional models (all $p < .002$).

The verb-clustering task was clearly more difficult for all models, featural and distributional, as evidenced by the much lower purity and higher entropy scores for the clustering solutions as compared with the nouns analysis. This finding is not surprising given that there were many more verb categories than there were for the nouns. Here, a modest but significant difference emerged between the VV feature representations and the distributional models.

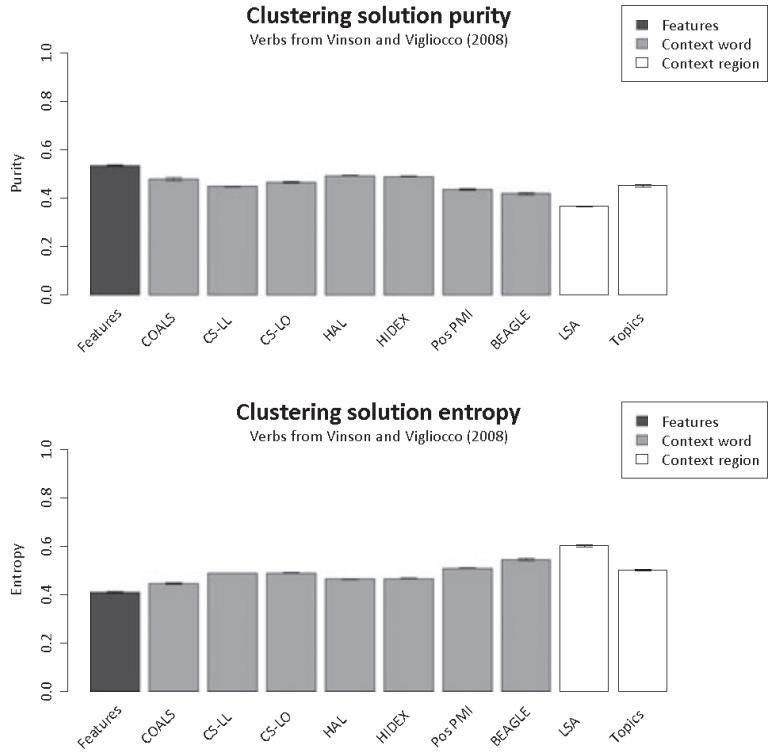

Fig. 4. Clustering solution purity (top) and entropy (bottom) scores for the verbs from Vinson and Vigliocco (2008).

5. Learning semantic representations from child-directed speech

In the analyses presented thus far, distributional models have shown a surprising ability to produce clusters of semantically related words corresponding to semantic classes of nouns and verbs. In some cases, these representations have demonstrated semantic structure that is similar to that found in feature-based representations derived from human data. In other cases, the clusters found from featural and distributional representations use different aspects of the data streams to arrive at their solutions.

Both the featural and distributional representations investigated in the previous two analyses are only designed to reflect the semantic knowledge of adults. However, a core question in cognitive science is how these representations are acquired. McRae et al. (2005) hypothesize that lexical representations result from repeated multisensory exposure to the referents of words. However, the details of how feature representations develop on this account have not been specified. The developmental account of distributional models, by contrast, is straightforward: Children may use the distributions of words in the language of their environment to develop and refine lexical semantic similarity relations. Thus, given a large corpus reflecting the linguistic environments of children, we can determine whether a distri-

butional model can learn psychologically plausible representations of semantic similarity. However, in order to assess such distributional model representations, comparable feature-based representations of word meanings are required.

5.1. The Howell et al. (2005) feature representations

Howell et al. (2005; hereafter HJB), modifying the methodology of McRae et al. (2005), derived child-relevant feature representations for words from adult feature-ratings. Their feature representations include both nouns (352) and verbs (90) with very early age-of-acquisition. HJB manually selected features related to children's sensorimotor experience, and then had adults rate how important each feature might be in children's representations of word meaning. As their focus was on prelinguistic knowledge, they concentrated on representing sensorimotor features of word meaning. In total, they coded for 96 noun features and 84 verb features with no overlap between noun and verb features. HJB had adults rate the relevance to young children of each feature; hence, each word was represented as a vector with 180 elements representing the average rating of the relevance of a given feature for that word. Examples of the feature representations for *bear* and *eat* are shown in Tables 2 and 6, respectively. Howell et al. (2005) demonstrated that using the feature representations in a recurrent neural network model of language acquisition led to improvements in word learning and word prediction.

5.2. Method

5.2.1. Semantic classes

Semantic classes for nouns were taken from the MacArthur-Bates Communicative Development Inventories (MCDI; Fenson et al., 1994). The MCDI is a tool widely used in language development research and contains norms for the age-of-acquisition of early-learned words. The MCDI divides nouns into a number of semantic classes. These classes represent

Table 9
MCDI semantic classes for nouns from the MCDI

Semantic Class	Number of Words	Example Words
Animals	39	Bird, zebra, bee, fish
Body parts	23	Leg, nose, tongue, tummy
Clothing	26	Shoe, belt, sweater, diaper
Food	58	Strawberry, carrot, hamburger, cereal
Furniture and Rooms	30	Drawer, tv, basement, bathroom
Outside Things	27	Moon, wind, tree, hose
People	26	Sister, teacher, babysitter, fireman
Places To Go	20	Playground, school, beach, outside
Small Household Items	44	Fork, telephone, pillow, plant
Toys	11	Puzzle, block, glue, story
Vehicles	12	Train, helicopter, bicycle, stroller

Table 10
VV semantic classes for action verbs from the MCDI

Semantic Class	Number of Words	Example Words
Body actions	21	Dance, play, eat, kick
Body sense	6	See, taste, hear, touch
Change of location	5	Bring, pull, push, put
Change of state	5	Dry, shake, spill, finish
Cognition	6	Hate, love, think, pick
Communication	6	Pretend, read, talk, show
Construction	5	Build, draw, fix, paint
Contact	7	Bump, catch, hit, open
Cooking	1	Cook
Destruction	2	Cut, rip
Exchange	3	Buy, give, take
Heat/light emission	0	
Motion direction	1	Go
Motion manner	8	Walk, swim, chase, stop
Noises	1	Sing
Noise-animal	0	
Tool action	1	Sweep

a somewhat different taxonomy than the WordNet semantic classes that is more relevant to children. For example, the MCDI contains classes such as Outside Things, Small Household Items, and Places To Go. The MCDI classes do not map exactly onto the WordNet classes. The Outside Things class includes some artifacts (*hose*), plants (*tree*), as well as entities from the natural world (*sun, moon*). The Places To Go class includes some artifacts (*backyard*), locations (*country, downtown*), and things that WordNet classifies as events (*party, picnic*). At the same time, the MCDI semantic classes are more fine-grained than those of WordNet, although both classifications cover roughly the same semantic space. In particular, the MCDI splits artifacts into a total of seven semantic classes: Clothing, Furniture and Rooms, Outside Things, Places To Go, Small Household Items, Toys, and Vehicles. We chose to use the MCDI semantic classes as a more appropriate structure to represent the semantic space of children. Table 9 lists the noun semantic classes and example words from each class.

The MCDI does not contain semantic classes for verbs. After some pilot work, we chose to use an extension of the VV verb semantic classes rather than the WordNet verb semantic classes. The WordNet classes for verbs are not always consistent. For example, *bite* is classified under Contact while *blow* is classified under Body (in the VV database these verbs are Body Actions). In other cases, the WordNet classes are not intuitive. An example is *clap,* which WordNet considers a verb of Communication, rather than a Body Action as in the VV database. To extend the VV classes to the MCDI verbs, an additional class, Cognition, was added to cover six verbs that did not fit well into the taxonomy (Table 10).

5.2.2. Distributional model implementation and training

To train the distributional models, a corpus of approximately two million words was constructed from caregiver speech in the American section of the CHILDES database (MacWhinney, 2000), representing input to children ages 12–48 months. This corpus was preprocessed in several ways. First, more than 700 word forms were standardized to eliminate orthographic variation. Second, all corpora were stemmed using a version of the Snowball stemmer (Porter & Boulton, 2006) augmented to change irregular verb forms to base forms. We chose to stem the corpus in order to increase the statistical foundation of content words in the corpus. We assumed that in child-directed speech, outside of sense-distinctions, the same word lemma is associated with the same meaning, regardless of morphological form (the same assumption we made about the TASA corpus). Word forms were not part-of-speech tagged because we do not assume children have this knowledge at the outset of word-learning. Third, most proper names in the corpora were converted to a single generic marker.

In the context word models, the context words were selected using an automatic procedure similar to that described in Section 3.2.3. For the CHILDES corpus, context words were allowed from any syntactic category. However, an extensive stop word list was employed in order to filter out very high-frequency and closed-class words. The stop word list contained 443 words with mostly grammatical functions: pronouns, determiners, auxiliary and light verbs, light nouns (e.g., *thing*), question words, high-frequency prepositions, conjunctions, quantifiers, and some adverbs. In addition, because of the nature of the corpus, the list contained contractions, hesitations (e.g., *uhuh)*, and some miscellaneous nonsense words. Finally, the context word selection procedure excluded proper names. The window size and shape for the context word models was the same as in the previous experiments.

Context region models require as input corpora that are split into contexts, such as documents. For example, the TASA corpus is split into context regions that average 151 words (Landauer & Dumais, 1997). Corpora from the CHILDES database, however, are not divided into coherent segments of interaction. For this investigation, we divided the corpus of caregiver speech into contiguous regions of 100, 500, and 1,000 words, and compared the clustering performance of LSA and the Topic model on each version. A factorial comparison of LSA models of several dimensionalities with the corpora split into different context regions revealed that the model trained on the 100-word context region corpus with 700 dimensions achieved the best clustering performance (see note 3). For the Topic model, a 100-topic model trained on the 500-word context region corpus produced superior clustering solutions.

5.2.3. Target words

The target words for this analysis were selected to meet several criteria. First, the words were shared between the HJB feature representations and the MCDI. Second, words had a frequency of 10 or more on the corpus derived from the CHILDES database that was used to train the distributional models (see Section 5.2.2 above). Third, in cases of noun-verb homonymy, we only included words that we judged to have a single dominant noun or verb

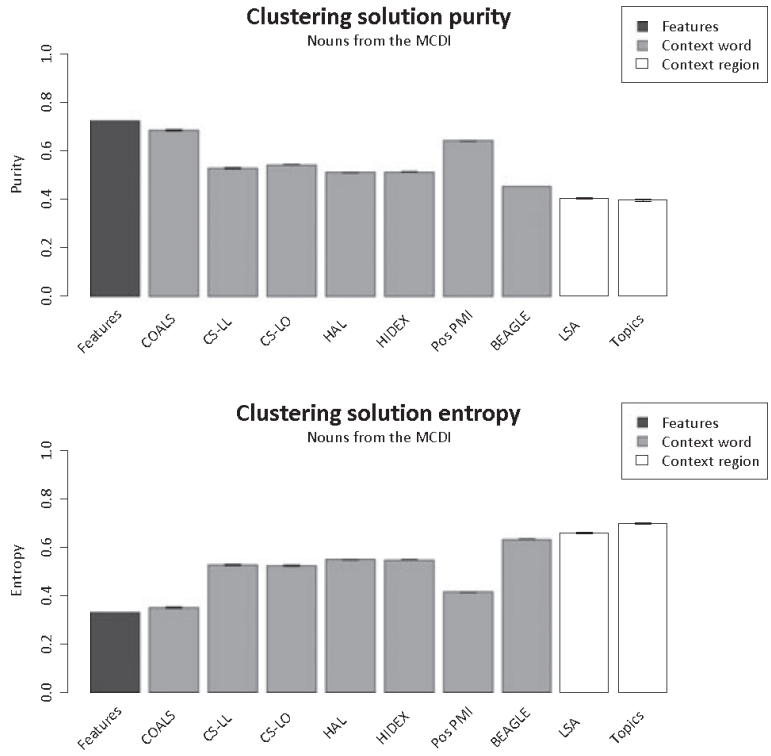

Fig. 5. Clustering solution purity (top) and entropy (bottom) scores for the nouns from the MCDI.

sense. Thus, we excluded words such as *clean* and *orange.* The target noun set consisted of 316 words and the target verb set contained 78 words.

5.2.4. Clustering

The same parameters were used to derive clustering solutions as in each of the previous analyses. For the noun dataset, we derived 11-cluster solutions, and for the verb dataset, 15-cluster solutions.

5.3. Results and discussion

5.3.1. Nouns

The purity and entropy scores of the clustering solutions for each model on the noun dataset are displayed in Fig. 5. Pairwise *t* tests with Bonferroni corrections on the purity scores showed that the HJB features ($M = 0.728$, $SD = 0.000$) were significantly higher than each of the distributional models (all $p < .0001$). The purity of COALS' solutions ($M = 0.689$, $SD = 0.013$) were significantly higher than all other distributional models (all $p < .0001$). The results for the entropy scores mirrored those of the purity scores. The HJB features ($M = 0.331$, $SD = 0.000$) were significantly lower than the distributional models (all

Table 11
Comparison of the clustering solutions of COALS and the HJB feature representations by semantic class for nouns from the MCDI (Note that cluster numbers begin with 0)

COALS				HJB			
Cluster	Semantic Class	Entropy	Purity	Cluster	Semantic Class	Entropy	Purity
6	Animals	0.186	0.895	0	Animals	0.000	1.000
				2	Animals	0.185	0.882
3	Body parts	0.239	0.852	9	Body parts	0.178	0.889
1	Clothing	0.139	0.920	5	Clothing	0.499	0.595
0	Food	0.289	0.524	3	Food	0.000	1.000
2	Food	0.118	0.935	4	Food	0.047	0.976
8	Furniture and Rooms	0.228	0.828	6	Furniture and Rooms	0.496	0.578
10	Outside Things	0.423	0.690	7	Outside Things	0.516	0.417
				10	Outside Things	0.509	0.464
4	People	0.754	0.368	1	People	0.248	0.839
7	People	0.545	0.385				
9	Small Household Items	0.478	0.613	8	Small Household Items	0.582	0.574
5	Vehicles	0.584	0.440				

$p < .0001$). COALS' entropy scores ($M = 0.350$, $SD = 0.016$) were lower than each of the other distributional models (all $p < .0001$).

As Fig. 5 shows, two context word models, COALS and Positive PMI, are able to achieve semantic structure comparable to the HJB feature representations simply by learning from children's ambient linguistic environment. Thus, we have evidence that the distributions of words even in child-directed speech are capable of producing representations with plausible similarity in semantic categories. In order to explore whether the trends in the types of representations derived by the feature and distributional models are similar to those for adults, we again conducted a detailed qualitative analysis, comparing the HJB features to COALS.

5.3.2. Comparison of HJB and COALS representations

Each model's best clustering solution in terms of purity and entropy was selected for analysis. The best HJB solution had a purity score of 0.728 and an entropy score of 0.331. The COALS solution had a purity score of 0.712 and an entropy score of 0.333. The majority semantic class label, entropy, and purity of each cluster are listed in Table 11. The HJB representations found clusters with majority labels for 8 of the 11 semantic classes; COALS found 9 of 11.

Both models found good clusters for several semantic classes, including Animals, Food, and Clothing. The HJB features split the animals into two clusters based on intuitive features such as <*has_eyes*> and <*breathes*> (Table 12): One mostly focused on flying animals such as birds and insects (cluster 2), and one including other animals (cluster 0). This result closely parallels the results of clustering the MCSM feature representations, suggesting a continuity of semantic representations for animals from early childhood on. COALS trained

Table 12
Comparison of the most descriptive and discriminating features for the clustering solutions of COALS and the HJB feature representations for the nouns from the MCDI (Note that cluster numbers begin with 0)

COALS						HJB					
Cluster	Semantic Class	Descriptive Feature	% Within-Cluster Similarity Explained	Discriminating Feature	% Between-Cluster Similarity Explained	Cluster	Semantic Class	Descriptive Feature	% Within Cluster Similarity Explained	Discriminating Feature	% Between-Cluster Similarity Explained
6	Animals	say	0.08	say	0.07	0	Animals	has_eyes	0.05	has_tail	0.07
		see	0.04	put	0.05			has_face	0.05	has_4_legs	0.07
						2	Animals	has_eyes	0.06	has_eyes	0.08
								breathes	0.06	breathes	0.07
3	Body parts	hurt	0.09	hurt	0.11	9	Body parts	is_solid	0.08	has_fur	0.13
		dry	0.05	dry	0.05			goodness	0.07	is_breakable	0.06
1	Clothing	wear	0.17	wear	0.21	5	Clothing	made_of_cloth	0.06	made_of_cloth	0.23
		put	0.09	put	0.03			goodness	0.06	is_comfortable	0.05
0	Food	drink	0.07	drink	0.09	3	Food	is_edible	0.12	is_edible	0.17
		spoon	0.06	spoon	0.07			is_delicious	0.11	is_liquid	0.17
2	Food	eat	0.17	eat	0.16	4	Food	is_edible	0.12	is_edible	0.27
		like	0.06	chocolate	0.03			is_delicious	0.08	is_delicious	0.19
8	Furniture and Rooms	kitchen	0.06	kitchen	0.10	6	Furniture and Rooms	is_solid	0.06	has_a_door	0.10
		put	0.05	sit	0.03			hardness	0.04	is_rectangular	0.07
10	Outside Things	see	0.04	sun	0.06	7	Outside Things	is_solid	0.09	made_of_stone	0.10
		sun	0.03	put	0.04			hardness	0.07	is_grey	0.06
						10	Outside Things	beauty	0.06	is_solid	0.08
								goodness	0.05	has_leaves	0.09
4	People	little	0.12	little	0.10	1	People	has_face	0.05	talks	0.07
		baby	0.05	baby	0.05			has_eyes	0.05	has_teeth	0.07
7	People	today	0.06	today	0.06						
		go	0.06	put	0.05						
9	Small Household Items	put	0.09	put	0.04	8	Small House-hold Items	is_solid	0.08	made_of_plastic	0.15
		need	0.05	need	0.04			hardness	0.05	made_of_metal	0.09
5	Vehicles	ride	0.09	ride	0.14						
		truck	0.04	truck	0.06						

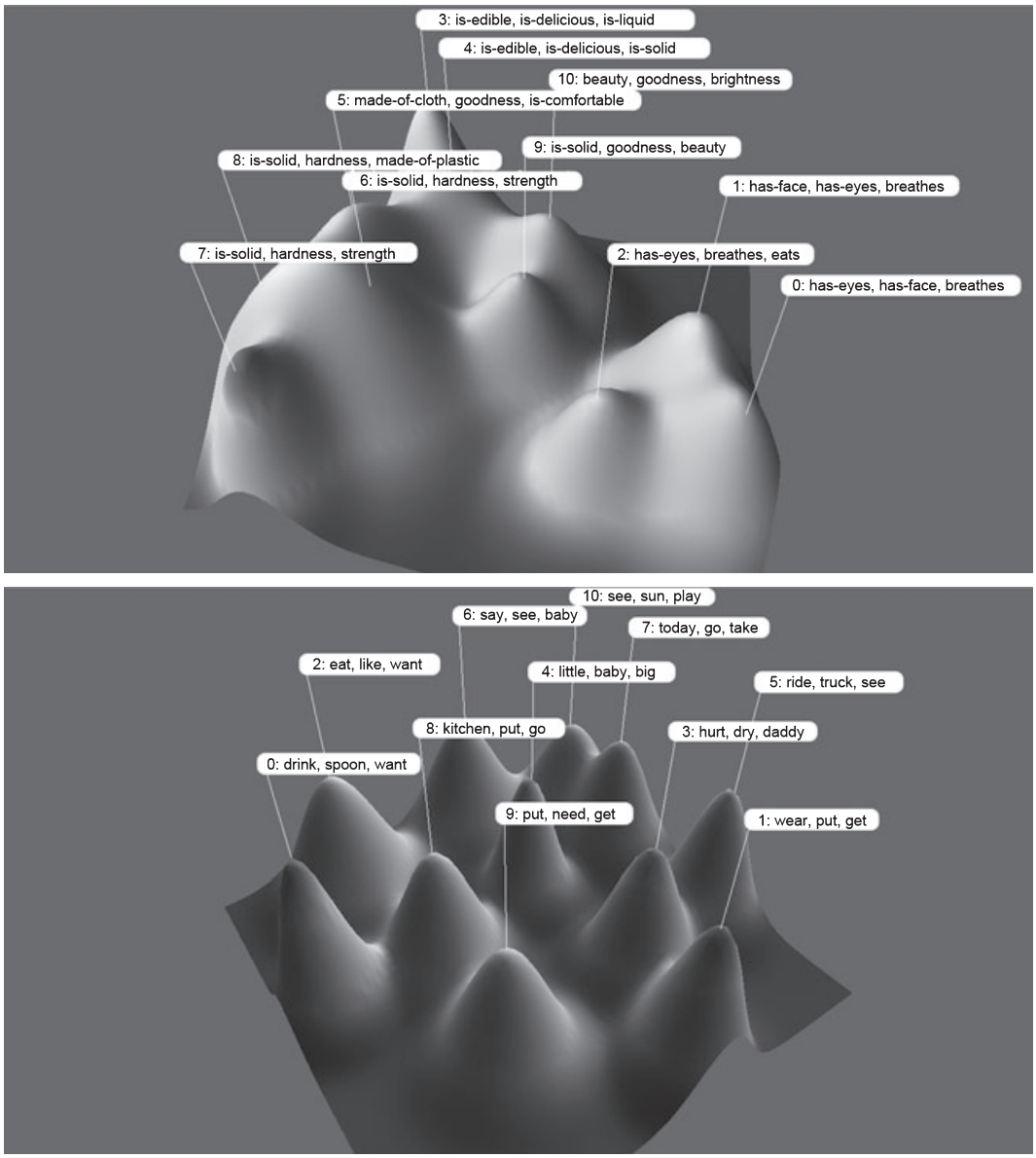

Fig. 6. Multidimensional scaling solution for an example 11-cluster solution from HJB (top panel) and COALS (bottom panel) for nouns from the MCDI. Peaks are labeled with cluster numbers and the three most descriptive features for each cluster.

on child-directed speech, like COALS trained on TASA, discovered only one Animals cluster, where similarity between animals was derived through the co-occurrence of words such *say* and *see*.

Both models found two clusters for Foods, one for liquid and one for solid foods. In these cases, the models used similar features: HJB's <*is_liquid*> feature was matched by COALS' *drink*, and COALS used *eat* while the HJB features used the feature <*is_edible*>. Hence, in the case of Food-related words, embodied interaction with the corresponding objects and actions is closely mirrored by the distributional statistics in linguistic experience.

Both models discovered a Clothing cluster, but the COALS cluster was much purer. Here, COALS found a similarity between words because of their co-occurrence with *wear* and *put* (as in *put on*). This result parallels the result for COALS trained on TASA. The HJB features lacked features specific to clothing and had to make use of less reliable features such as <*made_of_cloth*>. Thus, a salient part of children's experience with clothing—namely, the act of wearing—is reliably encoded in children's linguistic environments, but less so in human-generated sensorimotor-based features.

The multidimensional scaling solutions for each model are presented in Fig. 6. In the HJB solution (top panel), the ridge on the left is comprised of clusters that mainly contain artifacts (5, 6, 7, 8). The close proximity of these clusters stems from their shared descriptive features such as <*is_solid*> and <*hardness*>.

The COALS MDS solution (bottom panel) presents a strikingly different picture of children's semantic similarity structure for nouns. Although some related clusters appear

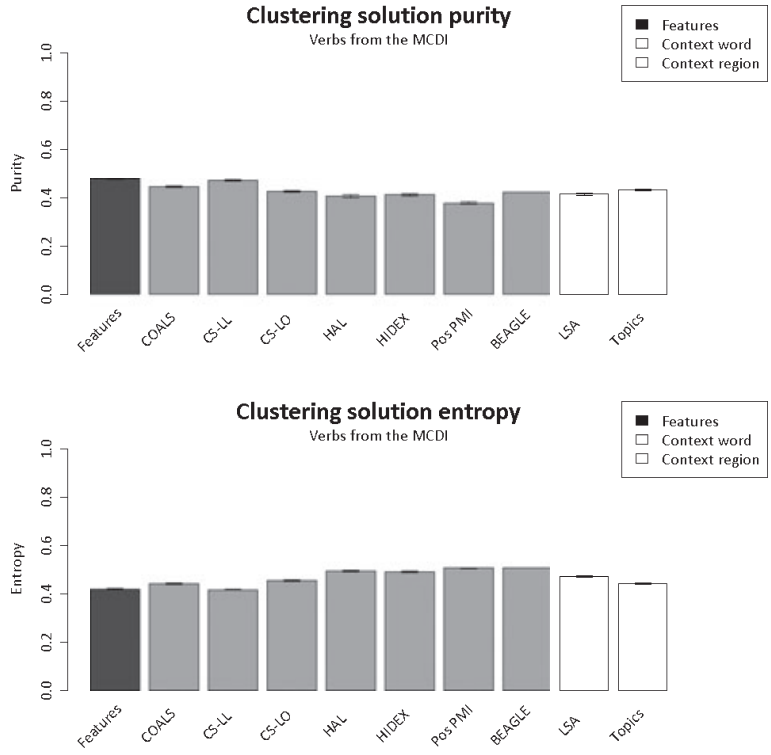

Fig. 7. Clustering solution purity (top) and entropy (bottom) scores for the verbs from the MCDI.

close together (e.g., the Food-related clusters on the left), almost all of the clusters form distinct peaks in the similarity space. This indicates that the COALS distributional representations are quite different between clusters. Thus, unlike the HJB representations, the distributional structure of child-directed speech provides disparate distributional profiles for different semantic classes of nouns.

Overall, the distributional features of the COALS model trained on child-directed speech were very different from the HJB features, and yet, just as in the MCSM/COALS comparison, the COALS representations were surprisingly effective for delineating clusters of semantically similar words. In a manner similar to COALS trained on TASA, COALS here emphasized the functions and behaviors of objects in its similarity structure. This produced some clusters very similar to those produced by the HJB features (e.g., the Food-related and Body Part clusters), while others were organized quite differently (Animals). The global similarity space of the COALS representations contrasted sharply with that of the HJB representations.

5.3.3. Verbs

Fig. 7 depicts the purity and entropy scores for each model on the children's verb dataset. Pairwise *t* tests with Bonferroni corrections on the purity scores revealed that the HJB features ($M = 0.478$, $SD = 0.006$) were significantly higher than each of the distributional models (all $p < .0001$) except CS-LL ($M = 0.472$, $SD = 0.010$, *ns*). CS-LL's purity scores were greater than each of the other distributional models (all $p < .0001$). CS-LL's entropy scores ($M = 0.417$, $SD = 0.009$) did not differ from those of the HJB representations ($M = 0.421$, $SD = 0.005$). These two models' entropy scores were significantly lower than the rest of the distributional models (all $p < .0001$).

On the verbs, then, the HJB features produce the best clustering solutions. Unlike in previous analyses, the best distributional model was not COALS, but CS-LL. However, it is clear from the figures that none of the models was able to cluster the verbs into semantic classes well. In this sense, the results are similar to those obtained in the analysis of the verbs from Vinson and Vigliocco (2008). In both cases, the semantic taxonomy was quite fine-grained (the number of semantic classes was 16 in the Vinson and Vigliocco analysis and 15 in this analysis). Thus, it is not clear whether the difficulty both types of models encountered was because of the meanings of the words or the semantic class distinctions. Indeed, because of the nature of the organization of verb meanings, which are generally thought to have a much ''flatter'' taxonomic structure (see e.g., Fellbaum, 1998), it may be difficult to tease these issues apart.

6. General discussion

We have reported analyses comparing two types of representations of semantic similarity: Representations that aggregate human-generated features and representations derived from simple statistical learning mechanisms operating on distributional information in the linguistic environment. The models were evaluated using a semantic clustering task, testing their abilities to recover similarity structure encoded in several existing semantic taxonomies.

Several distributional models were able to approach the clustering performance of the feature-based models on different datasets, but in all but one analysis the COALS model produced some of the best clustering solutions. Two aspects of COALS were particularly important for its behavior. First, COALS is a context word model: The model computes word-by-word co-occurrences within a moving window across the corpus. This type of model has been shown to be suited for recovering paradigmatic similarity relationships (Sahlgren, 2006), that is, words related by their occurrence in the same positions with respect to other words, for example, because they share the same syntactic category. Because the clustering task presented in this article used semantic classes that were restricted to nouns or verbs (no classes crossed syntactic category boundaries to group both nouns and verbs), the statistical learning mechanism of COALS may have been well suited to the task.

A second important difference between COALS and the other distributional models is its lexical association function. It appears that the lexical association function employed by COALS is particularly sensitive to covariation between words that is related to semantic similarity rather than chance co-occurrences. The clustering performance observed here suggests that a successful cognitive model should allocate attention towards stable distributional correlations when learning, and away from chance co-occurrences and anticorrelations. Further, the success of the COALS algorithm (at least in the realm of semantic categorization) constrains the set of plausible cognitive mechanisms that may be used by humans when encoding distributional information from the linguistic environment.

In recent work, Maki and Buchanan (2008) explored global similarities between measures of lexical semantic similarity. They compiled word–word similarities for concrete nouns and action verbs from feature representations (McRae et al., 2005; Vinson & Vigliocco, 2008), distributional models (LSA, the Topic model, and BEAGLE), association norms (Nelson, McEvoy, & Schreiber, 2004), and similarity computed from WordNet (Maki, McKinley, & Thompson, 2004). Maki and Buchanan demonstrated that these measures of semantic similarity were separable, encoding somewhat different types of semantic similarity, and they argued that distributional similarity is a distinct type of semantic similarity related to ''thematic'' associations between words. In the present work, in contrast, we employ a clustering evaluation, which has the advantage of abstracting away from idiosyncratic pairwise word similarities resulting from experimental task or corpus variation. Using this task, we demonstrated that feature-based and distributional models, despite being derived from different means, can produce similar clustering of semantically related words as measured by objective criteria. Importantly, this was most true of context word models. Hence, our results do not support the proposal that all distributional statistics reflect thematic similarity. Rather, different distributional similarity statistics are available in the linguistic environment, and some of these statistics—those computed by context word models—more closely reflect the similarity captured by feature-based models.

A recurrent theme in each of our analyses is that the streams of distributional and featural information are not completely redundant. Rather, they appear to be both correlated and partially complementary sources of information for semantic organization. Analysis of the clustering solutions demonstrated that distributional and feature-based models can, in many cases, find similar semantic clusters using the same ''features'' (e.g., the word *fly* in linguis-

tic experience, or the perceptual experience of seeing a bird fly). Much semantic information required for successful clustering is redundantly coded in the perceptual environment and in the distributional structure of the language that describes that environment. However, while redundant, each source often emphasizes different information that is useful for semantic clustering. For example, distributional models seem to give more weight to information about actions, functions, and situations, and less to information about direct perception related to objects (e.g., texture, internal properties, etc.). At least for context word distributional models such as COALS, this difference in emphasis may be the result of certain features being encoded less reliably in the linguistic environment as compared with the perceptual environment. Using the Wu and Barsalou taxonomy of feature types (see De Deyne & Storms, 2008), COALS was most able to learn situation-based features, such as the functions of objects and typical participants in events, taxonomic features, and the behaviors of entities. COALS, however, appeared to have much less access to other entity-related features such as external properties (e.g., strawberry <*is_red*>) and internal properties or materials (e.g., car <*made_of_metal*>).

In this sense, the two data streams provide complementary cues to semantic organization. While the feature-based models used information weighted heavily in that data stream, the distributional models made use of different cues to meaning that were salient in the linguistic environment but less well represented in the features. An ideal statistical learner could freely allocate attention to either source (or both) for the redundant information, but should extract information that is most available from the respective data streams to optimally organize semantic memory. During learning, children are likely to first glean information from sensorimotor experience. However, with acquisition of greater linguistic input, they may shift to extracting the redundant information from distributional structure and rely on perception for only the unique sources of information it provides. This proposition based on our data has much in common with Louwerse's (2007) symbol interdependency hypothesis.

It is important to note that the fact that the feature-based models often outperformed the distributional models in our analyses should not be taken as evidence that features are more important to semantic organization than is distributional structure. Such a conclusion falsely assumes that the models are equal in complexity and assumptions. The feature-based models were expected to outperform our distributional models because they use hand-coded representations based on human intuition, and this is a powerful wildcard in a model's performance (Hummel & Holyoak, 2003). Unlike classic feature models (e.g., Smith et al., 1974) all the feature representations used here aggregate human judgments across many individuals. However, they have human behavior encoded into them, in contrast with the distributional models which learn directly from linguistic structure (with the exception of our lemmatization/stemming). Hence, it is particularly impressive that a model like COALS learning from only raw data was able to rival—and in one case even outperform—the feature-based representations in semantic clustering.

Our results point to perceptual experience and linguistic distributional structure being both correlated and complementary data streams, and we believe that future modeling endeavors should be aimed at understanding the cognitive mechanisms by which the two sources are integrated to organize semantic memory. Each model type on its own can only

go so far—both sources are likely to be necessary to fully approximate the structure of human semantic memory. Recent attempts have been made to incorporate both sources into a single model of semantic representation (Andrews, Vigliocco, & Vinson, 2009; Durda, Buchanan, & Caron, 2009). Andrews et al. combine the representations of Vinson and Vigliocco (2008) with a distributional model similar to the Topic model. They demonstrate that the combination of representations provides better fits to semantic priming data. Finally, given the findings of this investigation and previous research that the different families of distributional models can produce quite different representational structure (Lavelli, Sebastiani, & Zanoli, 2004; Riordan & Jones, 2007; Sahlgren, 2006), future modeling efforts involving the integration of featural and distributional information should also explore different types of distributional information.

Notes

1. Our interest is not in the process by which lexicographers or semantic memory researchers assign words to semantic classes. Rather, we take these class assignments as a gold standard by which to evaluate the semantic clusters in the featural and distributional models.
2. In many implementations of HAL, 90% of the columns in the co-occurrence matrix with the lowest variability are discarded as a simple method of dimensionality reduction.
3. The best-performing LSA dimensionality depends both on the corpus on which a LSA is trained and on the semantic data it is intended to simulate, and it cannot be specified in advance. For example, Landauer and Dumais (1997), in simulating human performance on the TOEFL test, trained LSA on samples of the Grolier's Academic American Encyclopedia and explored a variety of dimensionalities. Among the dimensionalities they tested, several dimensionalities provided good fits to the data, with 300 dimensions performing best (see discussion in Quesada, 2007 and Dennis, 2007). For the MCSM dataset, we experimented with 100, 300, 500, and 700 dimensions. Ten clustering solutions were derived for LSA with each of the dimensionalities, then pairwise *t* tests with Bonferroni corrections were conducted on the purity and entropy scores of the clustering solutions, and the best performing dimensionality was selected. Note that the dimensionality used here differs from the 419 dimensions used on the LSA website (http://lsa.colorado.edu/).
4. We used the Topic Modeling Toolbox available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.
5. In this and the following analyses, we experimented with different numbers of topics for the Topic model. Twelve samples from the posterior distribution of words' probabilities for each topic were selected. Clustering solutions were derived for words' vector representations from each sample. Pairwise *t* tests using the Bonferroni correction were then conducted on the purity and entropy scores. For the MCSM dataset, simulations were run with 300 and 1,700 topics.

6. Because we take the semantic category labels in WordNet as a gold standard, we assume that the best models will produce closely matching clusters both in terms of number and composition. Requesting a larger number of clusters would improve the models' performance on the metrics introduced below (purity and entropy). However, these metrics only examine the composition of a clustering solution. The most rigorous test of the models is to examine not only whether they can produce clusters of similar makeup to the gold standard but also whether they can do this given the constraint to produce the same number of clusters.
7. Available at http://glaros.dtc.umn.edu/gkhome/views/cluto/.
8. Note that the variation in performance across clustering solutions for a given model is not psychologically meaningful; it is simply a result of randomization in the clustering algorithm.
9. The trends in the analyses reported here did not depend on the particular clustering solution that was selected. The 20 clustering solutions for each model showed virtually no variability across both the purity (MCSM $SD = 0.002505$; COALS $SD = 0.000000$) and entropy (MCSM $SD = 0.003959$; COALS $SD = 0.000671$) measures.
10. ''beh'' is an abbreviation of ''behavior.''
11. The *Noise-animal* class is not found in Vinson and Vigliocco (2008), but it is present in the data.

Acknowledgments

We thank Mike Gasser, Sandra Kübler, Chen Yu, Glenn Gunzelmann, and three anonymous reviewers, whose comments substantially improved this work. We are grateful to Steve Howell for providing his feature data. Mark Steyvers kindly made available additional code for running the Topic model. A preliminary version of the results reported in Section 5 was presented at the 2008 meeting of the Society for Computers in Psychology. This work was supported in part by Shared University Research grants from IBM, Inc. to Indiana University. BR was supported by NICHD postdoctoral training grant T32HD007475-13 and is now at Aptima Inc.

References

Andrews, M., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Barsalou, L. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Series B*, *358*, 1177–1187.

Barsalou, L., Santos, A., Simmons, K., & Wilson, C. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–283). New York: Oxford University Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

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.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, and Computers*, *30*(2), 188–198.

Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in a high-dimensional memory space. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual conference of the cognitive science society* (pp. 61–66). Mahwah, NJ: Erlbaum.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Connell, L., & Ramscar, M. (2001). Using distributional measures to model typicality in categorization. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference the cognitive science society* (pp. 226–231). Mahwah, NJ: Erlbaum.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science: A Multidisciplinary Journal*, *23*(3), 371–414.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. Klavans & P. Resnik (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (pp. 49–66). Cambridge, MA: MIT Press.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*, 213–231.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*(2), 145–193.

Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 57–70). Mahwah, NJ: Erlbaum.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Durda, K., & Buchanan, L. (2008). WINDSORS: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, *40*(3), 705–712.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods*, *41*, 1210–1223.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fenson, L., Dale, P., Reznick, S., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C. B. & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i185.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In Philological Society (Great Britain) (Ed.), *Studies in linguistic analysis* (pp. 1–32). Oxford, England: Blackwell.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). New York: Oxford University Press.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Glenberg, A., & Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in Neural Information Processing Systems*, *17*, 537–544.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Harris, Z. (1970). Distributional structure. In *Papers in structural and transformational linguistics* (pp. 775–794). Dordrecht, The Netherlands: D. Reidel Publishing Company.

Howell, S., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*(2), 258–276.

Hummell, J.E., & Holyoak, K.J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Jankowicz, D. (2005). *Modeling category-specific deficits using topographic, corpus-derived representations*. Unpublished PhD Thesis, McMaster University.

Johns, B. T., & Jones, M. N. (2009). Simulating false recall as an integration of semantic search and recognition. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2511–2516). Austin, TX: Cognitive Science Society.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

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.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Engelwood Cliff, NJ: Prentice Hall.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 1036). Mahwah, NJ: Erlbaum.

Karypis, G. (2003). *CLUTO: A Clustering Toolkit*. Available at: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download. Accessed April 15, 2008.

Kintsch, W. (2007). Meaning in context. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 89–105). Mahwah, NJ: Erlbaum.

Kintsch, W. (2008). Symbol systems and perceptual representations. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 145–163). New York: Oxford University Press.

Laham, D. (2000). *Automated content assessment of text using latent semantic analysis to simulate human cognition*. Unpublished PhD Thesis, University of Colorado.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lavelli, A., Sebastiani, F., & Zanoli, R. (2004). Distributional term representations: An experimental comparison. In D. Grossman (Ed.), *Proceedings of the thirteenth international conference on information and knowledge management (CIKM)* (pp. 615–624). New York: ACM Press.

Logan, G. D. (1988). Towards an instance theory of automization. *Psychological Review*, *95*(4), 492–527.

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. K. Landauer, D. S. Macnamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107–120). Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2008). Embodied relations are encoded by language. *Psychonomic Bulletin and Review*, *15*(4), 838–844.

Louwerse, M. M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively inspired NLP-based knowledge representations: Further explorations of Latent Semantic Analysis. *International Journal on Artificial Intelligence Tools*, *15*(6), 1021–1039.

Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 309–326). New York: Oxford University Press.

Louwerse, M. M., & Zwaan, R. A. (2009). Language encodes geographic information. *Cognitive Science*, *33*(1), 51–73.

Lowe, W. (2000). *Topographic maps of semantic space*. Unpublished Ph.D. thesis, University of Edinburgh.

Lowe, W. (2001). Towards a theory of semantic space. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd conference of the cognitive science society* (pp. 576–581). Mahwah, NJ: Erlbaum.

Lowe, W., & McDonald, S. (2000). The direct route: Mediated priming in semantic space. In L. Glietman & A. Joshi (Eds.), *Proceedings of the 22nd conference of the cognitive science society* (pp. 806–811). Mahwah, NJ: Erlbaum.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instrumentation, and Computers*, *28*, 203–208.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Behavior Research Methods*, *15*(3), 598–603.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, *36*(3), 421–431.

Martin, D., & Berry, M. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. Macnamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35–55). Mahwah, NJ: Erlbaum.

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Unpublished PhD Thesis, University of Edinburgh.

McDonald, S., & Lowe, W. (1998). Modelling functional priming and the associative boost. In S. J. Derry & M. A. Gernsbacher (Eds.), *Proceedings of the 20th annual conference of the cognitive science society* (pp. 675–680). Mahwah, NJ: Erlbaum.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

McRae, K., Cree, G. S., Westmacott, R., & de Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology*, *53*(4), 360–373.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 23–46). Cambridge, MA: MIT Press.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Perfetti, C. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, *25*, 363–377.

Porter, M., & Boulton, R. (2006). Snowball stemmer [computer software]. Available at: http://snowball.tartarus.org.

Pulvermüller, F. (2008). Grounding language in the brain. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 85–116). New York: Oxford University Press.

Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 71–85). Mahwah, NJ: Erlbaum.

Rao, V. A., & Howard, M. W. (2008). Retrieved context and the discovery of semantic structure. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20*. Cambridge, MA: MIT Press.

Rasmussen, M., Newman, M., & Karypis, G. (2003). *gCLUTO Documentation, Version 1.2*. Minneapolis, MN: University of Minnesota.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods*, *41*, 657–663.

Riordan, B., & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th conference of the Cognitive Science Society* (pp. 599–604). Austin, TX: Cognitive Science Society.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector Spaces*. Unpublished PhD Thesis, Stockholm University.

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1300–1305). Austin, TX: Cognitive Science Society.

Sanford, A. J. (2006). Semantics in psychology. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed. Vol. 11, pp. 152–158). Amsterdam: Elsevier.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. B. Jones & H. Somers (Eds.), *Proceedings of the international conference on new methods in language processing* (pp. 44–49). London: Routledge.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241.

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Mahwah, NJ: Erlbaum.

de Vega, M., Graesser, A., & Glenberg, A. (2008). Reflecting on the debate. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 397–440). New York: Oxford University Press.

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition*, *85*(3), B61–B69.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488.

Vigliocco, G., Vinson, D. P., & Siri, S. (2005). Semantic and grammatical class effects in naming actions. *Cognition*, *94*, B91–100.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183–190.

Vinson, D. P., Vigliocco, G., Cappa, S., & Siri, S. (2003). The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Understanding Language*, *86*(3), 347–365.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (Eds.) (1995). *The Educator's Word Frequency Guide*. Brewster, NY: Touchstone Applied Science Associates.

Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis (University of Minnesota Computer Science Technical Report CS 01-40)*. Minneapolis, MN: University of Minnesota.

Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In C. Nicholas (Ed.), *Proceedings of the eleventh international conference of information and knowledge management (CIKM)* (pp. 515–524). MacLean, VA: ACM Press.