Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Memory and Language 53 (2005) 258–276

Journal of Memory and Language

www.elsevier.com/locate/jml

A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning

Steve R. Howell [*], Damian Jankowicz, Suzanna Becker

McMaster University, Hamilton, Ont., Canada

Received 1 December 2004; revision received 15 March 2005
Available online 25 April 2005

Abstract

It is generally accepted that children have sensorimotor mental representations for concepts even before they learn the words for those concepts. We argue that these prelinguistic and embodied concepts direct and ground word learning, such that early concepts provide scaffolding by which later word learning, and even grammar learning, is enabled and facilitated. We gathered numerical ratings of the sensorimotor features of many early words (352 nouns, 90 verbs) using adult human participants. We analyzed the ratings to demonstrate their ability to capture the embodied *meaning* of the underlying concepts. Then using a simulation experiment we demonstrated that with language corpora of sufficient complexity, neural network (SRN) models with sensorimotor features perform significantly better than models without features, as evidenced by their ability to perform word prediction, an aspect of grammar. We also discuss the possibility of indirect acquisition of grounded meaning through "propagation of grounding" for novel words in these networks.

Keywords: Language acquisition; Features; Semantics; SRN; Neural network; Sensorimotor; Conceptual learning

Considerable evidence suggests that by the time children first begin to learn words around the age of 10–12 months, they have already acquired a fair amount of sensorimotor (sensory/perceptual and motor/physical) knowledge about the environment (e.g., Lakoff, 1987; Lakoff & Johnson, 1999; Bloom, 2000; Langer, 2001), especially about objects and their physical and perceptual properties. By this age children are generally able to manipulate objects, navigate around their environment, and attend to salient features of the world, including parental gaze and other cues important for word learning (Bloom, 2000). Some have suggested that this pre-linguistic conceptual knowledge has a considerable effect on the processes of language acquisition (Lakoff, 1987; Mandler, 1992; Smith & Jones, 1993) and even on later language processing (e.g., Glenberg & Kaschak, 2002; Barsalou, 1999). We also argue that the evidence indicates that this early prelinguistic knowledge has great impact, directly and indirectly, throughout a number of phases of language learning, and we attempt to begin to demonstrate this with a neural network model.

[*] Corresponding author. Present address: Department of Psychology (WJ Brogden Hall), University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53703, USA. Fax: +1 608 262 4029.
E-mail address: showell@wisc.edu (S.R. Howell).

0749-596X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2005.03.002

To begin with, this prelinguistic conceptual information helps children to learn their first words, which correspond to the most salient and imageable (Gillette, Gleitman, Gleitman, & Lederer, 1999) objects and actions in their environment, the ones they have the most experience with physically and perceptually. Generally speaking, the more "concrete" or "imageable" a word, the earlier it will be learned. This helps to explain the preponderance of nouns in children's early vocabularies (see Gentner, 1982). The meanings of verbs are simply more difficult to infer from context, as discussed by as demonstrated by Gillette et al. (1999). Only the most clearly observable or "concrete" verbs make it into children's early vocabularies. However, later verbs are acquired through the assistance of earlier-learned nouns. If a language learner hears a simple sentence describing a real-world situation, such as a dog chasing a cat, and already knows the words *dog* and *cat*, the only remaining word must be describing the event, especially if the learner already has built up a pre-linguistic concept of "dogs chasing cats" at the purely observational level. As Bloom (2000) describes, the best evidence for "fast-mapping" or one-shot learning of words in children comes from similar situations in which only one word in an utterance is unknown, and it has a clear, previously unknown, physical referent present. Of course, since the verb *chase* refers to an event rather than an object, the above example is not an exact fit to the fast-mapping phenomenon as it is usually described, but it is similar.

These very first words that children learn thus help constrain the under-determined associations between the words children hear and the objects and events in their environment, and help children to successfully map new words to their proper referents. This happens through the use of cognitive heuristics such as the idea that a given object has one and only one name (Markman & Wachtel, 1988), or more basic object-concept primitives (Bloom, 2000) such as object constancy.With a critical mass of some 50 words, children begin to learn *how to learn* new words, using heuristics such as the count-noun frame, or the adjective frame (Smith, 1999). These frames are consistent sentence formats often used by care-givers that enable accurate inference on the part of the child as to the meaning of the framed word, e.g., "This is a ___." These factors combine to produce a large increase in children's lexical learning at around 20 months. As they begin to reach another critical mass of words in their lexicon (approaching 300 words), they start to put words together with other words—the beginnings of expressive grammar (Bates & Goodman, 1999). Around 28 months of age children enter a "grammar burst" in which they rapidly acquire more knowledge of the syntax and grammar of their language, and continue to approach mature performance over the next few years.

By this account of language acquisition, conceptual development has primacy; it sets the foundation for the language learning that will follow. Words are given meaning quite simply, by their associations to real-world, perceivable events. Words are directly *grounded* in embodied meaning, at least for the earliest words. Of course, it may not be just simple statistical associations between concepts and words in the environment; the child is an active learner, and processes like joint attention or theory of mind may greatly facilitate the learning of word to meaning mappings (Bloom, 2000).

Of course, it seems clear that the incredible word-learning rates displayed by older children (Bloom, 2000) indicate that words are also acquired by linguistic context, through their relations to other words. Children simply are learning so many new words each day that it seems impossible that they are being exposed to the referents of each new word directly. The meanings of these later words, and most of the more abstract, less imageable words we learn as adults, must clearly be acquired primarily by their relationships to other known words. It may in fact be true that these meanings can *only* be acquired indirectly, through relationships established to the meanings of other words.

Evidence for the indirect acquisition of meaning is not limited to the speed with which children learn words. The work of Landauer and colleagues (e.g., Landauer and Dumais, 1997; Landauer, Laham, & Foltz, 1998) provides perhaps the clearest demonstration that word "meanings" can be learned solely from word-to-word relationships (although see Burgess & Lund, 2000; for a different method called HAL). Landauer's Latent Semantic Analysis (LSA) technique takes a large corpus of text, such as a book or encyclopedia, and creates a matrix of co-occurrence statistics for words in relation to the paragraphs in which they occur. Applying singular-value decomposition to this matrix allows one to map the words into a high-dimensional space with dimensions ordered by significance. This high-dimensional representation is then reduced to a more manageable number of dimensions, usually 300 or so, by discarding the least significant dimensions. The resulting compressed meaning vectors have been used by Landauer et. al. in many human language tasks, such as multiple choice vocabulary tests, domain knowledge tests, or grading of student exams. In all these cases, the LSA representations demonstrated human-level performance.

While models based on these high-dimensional representations of meaning such as LSA and HAL perform well on real world tasks, using realistically sized vocabularies and natural human training corpora, they do have several drawbacks. First, they lack any consideration of syntax, since the words are treated as unordered collections (a 'bag of words'). Second, LSA and HAL meaning vectors lack any of the grounding in reality that comes naturally to a human language learner. Experiments by Glenberg and Robertson (2000) have shown that the LSA method does poorly at the kinds of reasoning in novel situations that are simple for human semantics to resolve, due largely to the embodied nature of human semantics.

So it seems that there are two sources of meaning, direct embodied experience, and indirect relations to other words. However, there is an infinite regress in the latter. If words are only ever defined in relation to other words, we can never extract meaning from the system. We would have only a recursive system of self-defined meaning, symbols chained to other symbols (similar to Searle's Chinese Room argument, Searle, 1980). To avoid this dilemma, at least some of the words in our vocabularies *must* be defined in terms of something external. In children, at least, the earliest words serve this role. They are defined by their mappings to pre-linguistic sensory and motor experience, as discussed above. They do not require other words to define their meaning. The most imageable words are thus directly grounded, while the less imageable and more abstract words that are encountered during later learning are more and more indirectly grounded. At some point, we argue, the adult semantic system begins to look much like the LSA or HAL high-dimensional meaning space, with our many abstract words (e.g., love, loyalty, etc.) defined by relations among words themselves. However, the mature human semantic system is superior to the high-dimensional models, since it can trace its meaning representations back to grounded, embodied meaning, however indirectly for abstract words.

Intuitively, this is something like trying to explain an abstract concept like "love" to a child by using concrete examples of scenes or situations that are associated with love. The abstract concept is never fully grounded in external reality, but it does inherit some meaning from the more concrete concepts to which it is related. Part of the concrete words' embodied, grounded, meaning becomes attached to the abstract words which are often linked with it in usage. By our account, the grounded meaning 'propagates' up through the syntactic links of the co-occurrence meaning network, from the simplest early words to the most abstract. Thus we have chosen to call this the "propagation of grounding" problem. We argue that this melding of direct, embodied, grounded meaning with high-dimensional, word co-occurrence meaning is a vital issue in understanding conceptual development, and hence language development. We believe it is essential to resolving the disputes between embodied meaning researchers and high-dimensional meaning researchers.

In previous work (Howell & Becker, 2000, 2001; Howell, Becker, & Jankowicz, 2001) we began developing what we consider to be a promising method for modeling children's language acquisition processes using neural networks. In this work, we continue this effort, emphasizing the inclusion of pre-linguistic sensorimotor features that will ground in real-world meaning the words that the network will learn. This is a necessary precursor to addressing the "propagation of grounding" problem itself.

Our overall goal is to capture with one model the essence of the process by which children learn their first words *and* their first syntax or grammar. As mentioned above, this is a period stretching from the earliest onset of the first true words (10–12 months), through the "lexical-development burst" around 20 months up to the so-called "grammar burst" around 28 months. Developing a network that attempts to model the language acquisition that is happening during this period in children is, of course, an ambitious undertaking, and our models are still relatively simple. However, given the discussion on propagation of grounding above, this sort of developmental progression may actually be *necessary* not just for children learning language, but also for any abstract language learner such as a neural network or other computational model. That is, a multi-stage process of constrained development may be necessary to simplify the problem and make it learnable, with each 'stage' providing the necessary foundation for the next, and ensuring that meaning continues to be incorporated in the process. As such, we seek to develop and extend a single model that can progress through these 'stages' of language acquisition, from initial lexical learning, through rapid lexical expansion, to the learning of the earliest syntax of short utterances. Developing a model that fits developmental behavioral data on child language acquisition is one way to ensure that this process is being followed. For the simulations reported here, we have adopted and extended the Simple-Recurrent Network architecture that has been shown many times to be capable of learning simple aspects of grammar, namely basic syntax (e.g., Elman, 1990, 1993; Howell & Becker, 2001). Furthermore, SRN's have been shown to be able to produce similar results to high-dimensional meaning models. Burgess and Lund (2000) point out that their HAL method using their smallest text window produces similar results in word meaning clustering to an Elman SRN. Also, they state that the SRN is somewhat more sensitive to grammatical nuances. SRN's may be able to model the acquisition of meaning *and* grammar, unlike the high-dimensional approaches.

The present emphasis of our model is on the inclusion of sensorimotor knowledge of concepts or words (for clarity, in what follows we use the term "concept" to mean the mental representation of a thing or action, and the term "word" to mean merely the linguistic symbol that represents it). This pre-linguistic sensorimotor knowledge (following Lakoff, 1987) is represented by a set of features for each word presented to the network, features that attempt to capture perceptual and motor aspects of a concept, such as "size," or "hardness," or "has feathers". If a word that the network experiences is accompanied by a set of values or ratings on these feature dimensions, then the network should be able to do more than just manipulate the abstract linguistic symbol of the concept (the word itself). Like a child learning the first words, it should then have some access to the *meaning* of the concept. The network's understanding would be grounded in embodied meaning, at least at the somewhat abstracted level available to a model without any actual sensory abilities of its own.

Unlike most existing language models that employ semantic features (e.g., Hinton & Shallice, 1991; McRae, de Sa, & Seidenberg, 1997) our sensorimotor feature set has been designed to be pre-linguistic in nature. That is, features that derive from associative knowledge about which words occur together or other language-related associations are excluded. Only features that a preverbal child could reasonably be expected to experience directly through his or her perceptual and motor interactions with the world are included. As discussed above, while children's first words are obviously learned without any knowledge of language-related word associations, children quickly begin to incorporate linguistic associative information into the semantic meanings of concepts. Certainly, at some point words begin to acquire meaning not only from the sensory properties of the concept, but from the linguistic contexts in which the word has been experienced. We take the conservative stance herein of excluding any linguistic associative influences on sensorimotor meaning; the sensorimotor feature representations do not change with linguistic experience. This is primarily for practical issues of implementation. The network is capable of learning these associations, but they do not affect the sensorimotor feature representations directly.

Whereas most language models employ binary features, our features are scalar-valued (range 0–1), allowing a network to make finer discriminations than merely the binary presence or absence of a feature. For example, two similar items (for example, two cats) may be perceived, but they are not identical; one is larger. Our dimension of size would differentiate the two, with one receiving a rating of 0.2, one of 0.3. Binary features cannot easily make such fine distinctions. Finally, inspired by the work of McRae et al. (1997) on human-generated semantic features, the feature ratings that we use are all derived empirically from human participants.

One of the advantages of the neural network model of child language development that we present below is the ability to measure word-learning performance using analogues of lexical comprehension tasks that have been used with children. Since the network learns to associate the sensorimotor features of each concept with a separate phonemic representation of the word, it is possible to examine the strength of the connection in either direction. Thus, given the phonemes of the word, we can measure the degree to which the network produces the appropriate sensorimotor meaning vector. This we refer to as the 'grounding' task, analogous to when a child is asked questions about a concept and must answer with featural information, such as "What kind of noise does a dog make?" or "Is the dog furry?" Similarly, we can also ask if, when given the meaning vector alone, the network will produce the proper word. This is an analogue to the 'naming' task in children, where a parent points to an object and asks "What is that?" In the network, if the completely correct answer is not produced, we can still measure how close the output was to the correct answer. For example, we can check whether the answer was a case of "cat" produced in place of 'dog', two concepts with a high degree of featural overlap, or whether it was a complete miss. In this paper, we address the grounding task, but not the naming task, although the model can account for both. However, the central aim of this paper is to investigate the contribution of the sensorimotor features to improving the model's lexical and grammatical learning.

In Experiments 1 and 2, we describe the empirical collection of feature ratings for nouns and verbs, respectively, and describe the results of several analyses performed to verify that they are capturing an abstract representation of the words' meanings. In Experiment 3 we describe simulations of a neural network model using these features and trained with a large naturalistic corpus of child-directed speech. We examine the extent to which the inclusion of sensorimotor features improves lexical and grammatical learning over a control condition, in an attempt to demonstrate the utility of feature grounding for language acquisition. However, it is important to note that in referring to "grammatical learning" we are in fact only considering the simplest aspects of grammar, namely basic sequence learning.

Experiment 1—Generation of noun sensorimotor features

Developing a set of sensorimotor dimensions that are plausible for 8- to 28-month-old infants was an important first stage of this research effort. In our previous models of lexical grounding and acquisition of grammar (Howell & Becker, 2001; Howell et al., 2001), we used a more simplistic semantic feature representation of words (Hinton & Shallice, 1991) that was both artificial and confounded words' conceptual semantics with "associative semantics," the linguistic relationships between words. We needed a more child-appropriate set of semantic features. Of course, developing these semantic features actually involves two issues: one, collecting the ratings, and two, making sure that the results of the ratings actually reflect realistic early semantic relationships. If the collected ratings do not have plausible semantic cohesion, they cannot be very useful in further work.

Method

To avoid having artificial, experimenter-created semantic feature representations, we investigated the McRae et al. (1997) empirically generated feature set. However, of the thousands of features contained in that set, many were non-perceptual (e.g., linguistically associative), and few were common across many concepts. To obtain an appropriate set of input features for a neural network model of child language acquisition, we required a more compact, concrete set of features that are perceptual and motor in nature, and could reasonably capture purely pre-linguistic knowledge. Thus, we narrowed down the McRae et al. feature list to some 200 common and widely represented features. This list was further condensed by converting each set of polar opposites and intermediate points to a single set of 19 polar-opposite dimensions. For example, "small" and "large" became a single continuous dimension of size, ranging from small (0) to large (10), and eliminating the need for "tiny," "medium," "huge," etc. The remaining 78 features which could not be unambiguously converted to a set of polar opposites were retained as a condensed list of scalar-valued dimensions, such as color (is_red) or texture (has_feathers), where the numeric value indicated the probability of possession of that feature by that concept. We use the term 'feature dimension' or 'dimension' to refer to all 97 dimensions, however, since when considered as components of a meaning vector they each represent a spatial dimension in a 97-dimensional space.

This resulting list of features was then reviewed by an independent developmental psychologist, for accessibility to children of the age range in question (8–28 months), and any features that were not considered developmentally appropriate were removed. For example, "age" is not reliably perceived by children beyond simply "young" or "old" (Dr. Laurel Trainor, private communication, 2001) and so was removed.

The final list of 97 sensorimotor feature dimensions (see Appendix A) was small enough to be feasible as input for our neural network models, and broad enough to be applicable to many concepts. Given this set of feature dimensions, it was next necessary to obtain ratings of the early concepts along each feature dimension. We used a large sample of human raters to generate the featural ratings for our early words. Our raters were undergraduates at McMaster University who participated in this experiment for course credit in an introductory psychology course.

Participants were presented with the concepts and the list of feature dimensions along which to rate them on a computer screen. The display was presented via a web browser, and responses were entered by filling in response boxes on the display. Participants were given detailed instructions (see Appendix A) as to how to make judgments, and which anchoring points to use in assigning numerical values. For example, in rating the size of an object, the smallest item a child might know about might be 'pea,' while for adults it might be something microscopic like 'virus.' Thus, participants were specifically instructed to make judgments taking into account the limited frame of reference that a pre-school child would have, especially relevant for polar-opposite dimensions such as "size." Participants entered their data as numbers between 1 and 10, which were later scaled down to the 0–1 range for easier presentation to neural network models.

The rating forms were administered over the Internet as web forms. The data was checked carefully for outliers. Three participants' data were excluded due to obvious response patterns (all 0's, all 10's, 1-2-3's, etc.), indicating insufficient attention given to the task. Ratings were collected for 352 noun concepts from the MacArthur Communicative Development Inventory (MCDI, Fenson et al., 2000) in 38 separate phases with approximately 10 concepts each during winter, 2002. The first two phases had 10 participants each; the rest had 5 participants each, for a total of 200 participants. Participants received course credit for participation so long as the data were not obviously invalid as discussed above The resulting ratings were then averaged across participants yielding a single feature vector of size 97 for each concept, 352 in all.

Three forms of analysis were performed on these newly created feature representations, in order to demonstrate that they do capture important aspects of the meanings of the words represented: a hierarchical cluster analysis, a Kohonen self-organizing map, and a Euclidian-distance-based categorical membership analysis.

Results

We analyzed the 352 averaged feature vectors in a hierarchical cluster analysis using SPSS version 11.5, to see whether our features captured our intuitive sense of word similarity. The 352 concepts clearly clustered by meaning, with subcategories merging nicely into superordinate categories (see Appendix B). Animals are separated from people, people and animals are separated from vehicles and inanimate objects, etc. Thus, while the high degree of variability between participants' ratings was originally a concern, after averaging, the regularity inherent in the feature vectors is quite reassuring. To provide another view on the ratings, the ratings vectors were fed into a Self-Organizing Map (Kohonen, 1982, 1995) neural network, which sought to group the concepts topographically onto a two-dimensional space based on their feature similarity. The resulting topographic organization respects the semantic similarities between concepts, showing intuitive groupings based only on the sensorimotor features of concepts (see Fig. 1). Note, for example, the grouping of "creatures that fly" in the top left corner, and the grouping of parts of the body in the middle-left.

A more clearly defined measure of success is provided by the categorical analysis. We formed category centroids for each of the pre-existing categories of nouns on the MCDI form from which the words were originally drawn. This was done by taking all of the words that belonged to that category and averaging together their feature vector. Then each and every word's feature vector was compared to the centroids of each of the 11 categories represented, and the closest match indicated into which category the word should fall. This was done both with the target word included in the centroid generation process, and with it excluded (a more conservative approach). Results are very good, at 92.8 and 88% accuracy, respectively (Chance performance would be 9.1%). See Table 1 for details.

Discussion

We believe all three analyses indicate the success of the experiment. The hierarchical clustering analysis,

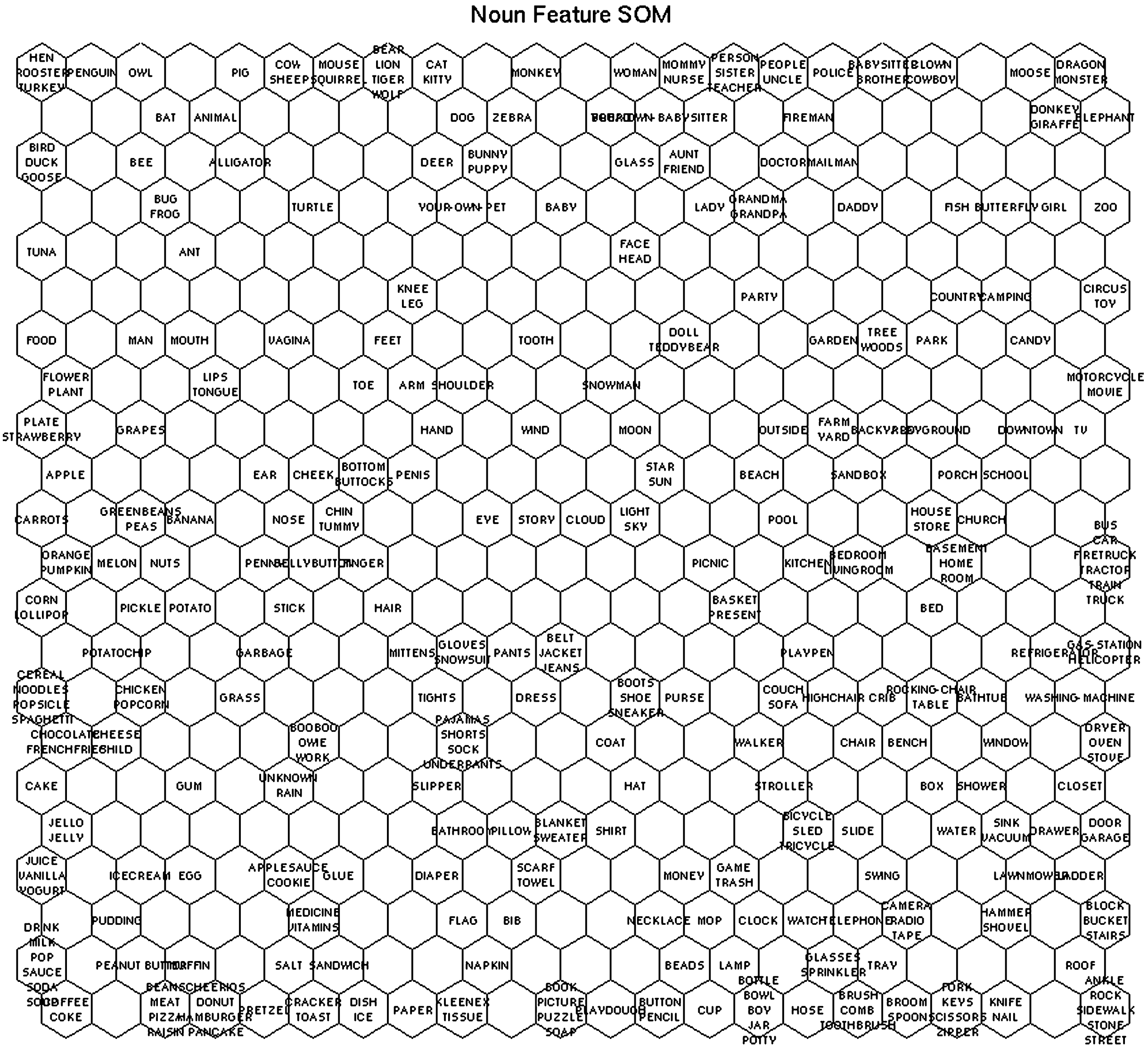


Fig. 1. Self-organizing feature map of Experiment 1 feature vectors. Each concept is written on the unit that responded most highly to presentation of that concept after training. Note the grouping of similar concepts on nearby units, as well as the overall topography of similarity.

Table 1
Noun category agreement results feature vectors compared to centroids of categories drawn from MCDI

Category number	Category name	Inclusive accuracy	Exclusive accuracy
1	Animals	0.8205128	0.8205128
2	Vehicles	0.9166667	0.75
3	Toys	0.9166667	0.8333333
4	Food and drink	1	1
5	Clothing	0.9285714	0.8928571
6	Body parts	1	0.862069
7	Small household items	1	1
8	Furniture and rooms	0.8484848	0.7878788
9	Outside things	0.9333333	0.8666667
10	Places to go	0.8636364	0.6363636
11	People	0.8461538	0.8461538
	Overall	0.9283668	0.8796562

while vast and somewhat difficult to interpret, shows many clear separations of concepts, and consistent local clusters of meaning. The SOM representation shows clear clustering by meaning, with both fine-grained and broader similarity structures across the map. Finally, the categorical analysis provides a clear numerical measure of the goodness of fit of our features to the preexisting categorizations of these nouns, with 93% accuracy of word to category. The sensorimotor feature ratings thus capture much of the meaning of the concepts, definitely enough to be useful as inputs to our language learning model, and they certainly capture what's important for categorical reasoning.

Experiment 2—Generation of verb sensorimotor features

In this experiment we followed much the same methodology as for Experiment 1, this time for verb features. However, given that verbs correspond to events in the world rather than to objects, the nature of verb features was expected to be different from that for nouns. Also, there was no pre-existing taxonomy of verb features readily accessible in the literature, as there had been for nouns (what we mean by verb features is different from verb 'classes,' the way verbs are usually grouped). Therefore, our collection of verb features proceeded in two steps. First we conducted a pilot experiment in verb feature generation with human participants, and from that we created a set of verb feature dimensions to be rated in a web-based phase of the experiment exactly as in Experiment 1.

Method

The pilot experiment was conducted with 12 undergraduate participants at McMaster University (see Appendix C for the instructions given to participants). Participants completed a feature generation form for some of the earliest (MCDI, Fenson et al., 2000), and most prototypical (Goldberg, 1999) verbs, with the objective being not complete characterization of any given verb but rather the creative generation of a set of feature dimensions which might be common to many verbs.

While fully half of the features generated were unusable due to contamination by functional relationships with corresponding nouns, associational relationships, etc., there were sufficiently many perceptual and motor features identified to allow us to create an initial set of feature dimensions. From this beginning, we were able to fill in missing complements of existing dimensions. For example, several participants focused on limb movement to define verbs, which is in line with some existing models of verb definition in computer science (see for example Bailey, Feldman, Narayanan, & Lakoff, 1997). From this and considerations of bodily motion and proprioceptive constraints in humans we were able to generate a large primary set of joint-motion dimensions. We also included some other features that had been identified by pilot participants, which brought the total to 84 feature dimensions (see Appendix C for a list)

A second group of 45 participants participated in the rating phase of the verb experiment (see Appendix C for the instructions given to participants). As in Experiment 1, they rated each verb on the list with a value between 0 and 10 on the 84 feature dimensions. Each participant rated 10 of the concepts. We then converted these ratings to the 0–1 range, which became the feature representations for verbs used in the Experiment below. We analyzed the results of the experiment (the feature ratings) in the same three ways as in Experiment 1: a hierarchical cluster analysis (see Appendix D), a self-organizing map (Kohonen, 1982, 1995), and a Euclidian-distance-based categorical membership analysis. The categories used in the latter analysis were drawn from Levin (1993) and grouped together into superordinate categories with the assistance of linguists Anna Dolinina of McMaster University, and Silvia Gennari of the University of Wisconsin – Madison. Nine categories were used, as can be seen in Table 2. Only the inclusive methodology was used to create the category centroids, based on the results from experiment 1.

Table 2
Verb category agreement results feature vectors compared to centroids of categories drawn from MCDI

Verb category	Percentage correct
Body-movements	60
Motion	83
Creation/destruction	64
Food-related	83
Possession and relocation	78
Change of state	55
Statives	78
Communicative	67
Perception	75
Overall	70

Results

Overall, the meanings of verbs do not cluster as coherently as do the meanings of nouns. Still, as can be seen from the SOM, similar verbs do group together in space (see Fig. 2). Note the grouping of "tongue-verbs" in the top left, and movement verbs in the bottom right, for example. Major trends in the cluster analysis for verbs are less clear than for nouns, although the analysis does find many intuitively reasonable group-

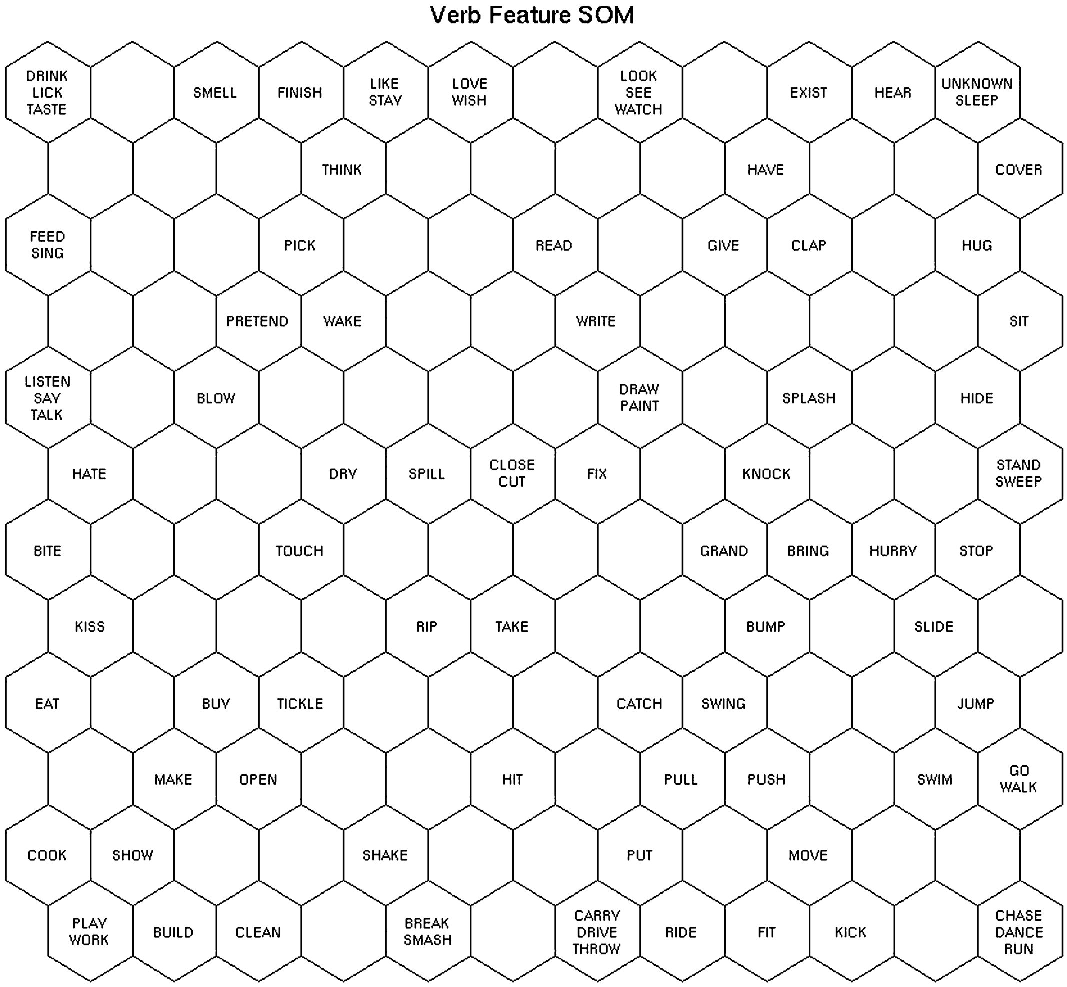


Fig. 2. Self-organizing map of the verb feature ratings. Note the grouping together of words involving similar motor activities such as drink/lick/taste and listen/say/talk as well as modes of locomotion such as slide/jump/go/walk/hurry.

ings, such as take, bring, push, put, and move, for example (see Appendix D).

Finally, the categorical agreement analysis, while not as clear as that for the nouns shown previously, still demonstrates a 70% overall accuracy of the target words to their correct category. Chance performance would be 11.1%. Categorical performance by category is shown in Table 2. The accuracies range across category, from 83% for motion or food-related categories, to a low of 55% for change of state verbs.

Discussion

The somewhat weaker clustering of our verb features is consistent with the results of Vinson and Vigliocco (2002), who also show that verbs generally do not cluster very well. Their verb features were also human-generated, but they placed no developmental or sensorimotor restrictions on the form of those features as we did in this experiment. Nonetheless, it seems that verbs, or pre-linguistic verb concepts, simply do not share as tight a similarity space as nouns do, although the fact that there were fewer verbs in Experiment 2 than there were nouns in Experiment 1 may have an effect, as there is less opportunity for featural similarity to become apparent. Also, when examined by category, the verb accuracies seem to be generally highest for the more concrete verb categories, and lowest for the more abstract (e.g., change of state). However, our features are still capturing important aspects of the meanings of verbs, as can be seen qualitatively in the hierarchical cluster analysis and SOM, and quantitatively in the Category Agreement analysis. An agreement rating of 70% is more than sufficient for us to wish to use these features in further experiments.

In Experiment 3, we investigated the contributions of sensorimotor feature grounding, in both nouns and verbs, to language learning in a neural network simulation.

Experiment 3—A large corpus model of grounded language acquisition

In previous work (Howell & Becker, 2001), we determined that adding an artificial set of semantic features to an SRN improved word prediction dramatically (18.5%–37.1%). However, in that experiment the word representations were localist (a series of zeroes with a single 1), while the feature representations were binary distributed codes (a sequence of zeros and ones). It was impossible to determine how much of the improvement in word prediction was due to the simple increase in the information content of the combined input representation, rather than the inter-word similarity structure inherent in the semantic features. In contrast, in this experiment the word representation is a very long (140 elements) distributed representation of phonemic features. The feature representations are smaller, 97-element (noun) or 84-element (verb) vectors of scalar-valued features. Also, as discussed below the control condition features are matched for numerical range and variability.

Additionally, in that previous model both the phonological information and the semantic information were presented as inputs to the network. Since we are arguing that children have these early semantic concepts pre-linguistically, it makes more sense to use the semantic features as output targets instead of inputs. Children have already formed internal representations of the features of a concept by the time of initial language learning. By using these features as output targets, rather than inputs, the network is forced to focus on the mapping of words to meanings and therefore learning the associations between words and existing concepts, as well as how those words predict each other in the speech stream.

In this experiment, then, any benefit from the inclusion of semantic information is thus expected to be due to the statistical regularities inherent in the sensorimotor feature information, beyond a simple increase in the information content due to the use of distributed representations. Specifically, we hypothesized that word prediction, a measure of syntactic learning which is one part of grammar (Elman, 1990), would improve with sensorimotor grounding of nouns and verbs. Essentially, meaningful semantics should improve syntactic learning.

Method

We modified the Simple-Recurrent Network (SRN) architecture to perform three separate tasks simultaneously, in three separate pools of output units (see Fig. 3). A small common hidden layer and context layer of 10 units each were used, to force the network to develop an integrated internal representation common to the three tasks. This may, in fact, be a good analogue to children's early learning, when attentional and other resources are immature and very limited (Newport, 1990). A single input layer presented whole-word phonetic representations of words in serial order through the corpus. Each word was encoded as a set of up to 10 phonemes using 140 input units. The 140-element word inputs represented 10 phonemic slots each of 14 phonemic feature bits, without representation of word boundaries. The Carnegie Mellon University (CMU) machine-readable phonetic transcription system and pronouncing dictionary was used to generate our phonetic representations of words (available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Each phoneme was uniquely mapped to a set of 14 bits

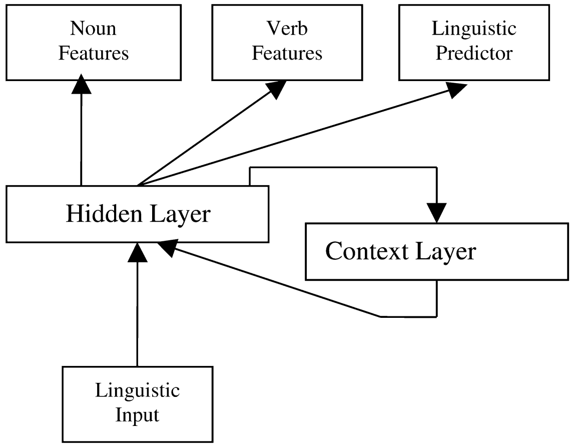


Fig. 3. Modified SRN architecture, including standard SRN hidden layer and context layer, standard linguistic (word) prediction output, and novel noun feature output and verb feature output. The linguistic input is a whole-word phonetic representation of up to 10 phonemes. The Noun and Verb feature targets are meant to be an abstract representation of pre-linguistic sensory and motor-affordance semantics.

representing articulatory dimensions of the phonemes. Words shorter than 10 phonemes had their rightmost slots padded with 14 zeros, while longer words were truncated.

The Linguistic Predictor output layer performed the word prediction task: predicting from the current input word what the next word would be. At each time step, its task was to predict the phonemic representation of the input word at the next time step. The task for the remaining outputs was to produce the sensorimotor features of the current word. The Noun Features layer had output targets that represented the sensorimotor features for the current word, as created in Experiment 1. The Verb Features layer had output targets that represented the sensorimotor features for the current word, as created in Experiment 2. When the current input was not a noun or a verb (respectively), a vector input of all 0's was presented at that layer, and no backpropagation of error was performed for that layer.

Employing the sensorimotor features as output targets was partly designed to eliminate the confound of representational richness involved in using additional inputs, as discussed above. Also, the fact that the network is producing sensorimotor noun and verb features at the output means that we can examine the ability of the network to generate the correct features for any given word. This gives us a measure of vocabulary acquisition, or lexical learning, both during learning and when testing generalization performance on novel words presented at the input.

Corpora and training schedule

We used a large (8328 word) selection of speech drawn from the CHILDES database (MacWhinney, 2000) transcribed from mother–child playtime interactions. This corpus was created by appending all of the Bates FREE20 data sets (Bates, Bretherton, & Snyder, 1988; Carlson-Luden, 1979) from the CHILDES database into a single body of text without pauses or sentence markers.

Two conditions of the network were run to simulate an experimental condition and a control condition. The Experimental condition used the full network as described above. The Random Control network used the same architecture as the Experimental condition, but replaced the human-generated (and meaningful) semantic features with randomized permutations of that same set of features. This condition is intended to control for sheer number of connections and input vector magnitudes. The randomization was performed by iteratively swapping the value at each position on the 97 element vector with that of another random position. When all words' representations had been randomized, each word's entire randomized feature representation was then swapped with another word's representation. This manipulation minimizes any featural similarity between related words.

Ten networks were run in each condition, for a total of 20. Each network was run for 200 epochs using the SRNEngine simulation package (Howell & Becker, 2005). Training used the back-propagation of error learning algorithm (Rumelhart, Hinton, & Williams, 1986). Rather than running these large networks to asymptotic performance, we simply ran them for a fixed period (200 epochs) within which grammatical prediction began to approach reasonable levels of performance. Due to the computational demands of the process, the networks' word prediction (grammatical) accuracy was calculated and recorded only at 50 epoch intervals. The network used a Euclidian-distance-based output rule to convert its output activations to a word label; thus every time step resulted in a discrete word prediction, as opposed to any sort of phonological blend state. Comparison of this word to the target word produced the accuracy measure.

This 'exact prediction match' criterion is quite conservative. The predicted word has to be the exact target word expected, or it is incorrect. Thus, we also used a second accuracy measure, a more generous (and arguable more accurate as a measure of grammatical learning) "categorical match" criterion, where the predicted word only had to be in the same grammatical category as the target word. All words in the corpus were divided into 1 of 12 grammatical categories, which included: adjective, adverb, conjunction, determiner, other, noun, possessive, preposition, pronoun, meaningless, and verb. The inclusion of this measure is to guard against the possibility that our exact match criterion is too conservative to have enough power to detect a difference between the Experimental and Control Conditions.

Also, as an analogue to lexical learning, we analyzed the data from the other two output layers, the Noun Feature encoding accuracy and the Verb Feature encoding accuracy, to see if there was any difference in the accuracy between Experimental and Control conditions, and if there was any relationship to the frequency of the word in the corpus.

Results

The results show a small but significant difference in word prediction accuracy (7.5% vs. 8.6%) between the two conditions (see Fig. 4). Using the exact match error criterion, the difference between the two conditions at epoch 200 is significant (*t* test at epoch 200, $p = .017$, $df = 18$). The percentage difference between the two conditions is 13%. Also, the gap between the two conditions is wider in the later epochs than in the earlier ones. Indeed, a repeated measures ANOVA on the data from epochs 150 and 200 yields a significant interaction effect of training by condition ($p = .034$, $df = 18$).

Using the categorical match error criterion, the mean accuracy of the Experimental group rises to 0.185, the control group to 0.171. The size of the difference is 0.014, or an 8.2% difference between the two groups. The difference under this error criterion is also significant, (*t* test at epoch 200, $p = .035$, $df = 18$). Due to the processing demands of calculating this error criterion, it was only calculated for the final epoch of training.

Noun encoding accuracy is also significantly different (approximately 11% difference) between the two conditions after 200 epochs (*t* test at 200 epochs, $p = .0344$, $df = 18$), with the sensorimotor feature condition being superior to the random features condition (see Fig. 5). The difference in verb encoding accuracy was not significant, however ($p = .120$, $df = 18$).

To examine further the trajectory of performance, we ran one of the Experimental condition networks above (chosen at random) for a total of 500 epochs. At this point, noun and verb grounding were quite good, as can be seen from Table 3 below, although based on past experience accuracy could rise as high as 90% with further training. The network did not learn to accurately produce sensorimotor features for any noun that occurred fewer than 4 times in the corpus, nor for any verb that occurred fewer than 5 times. Feature production accuracy for both nouns and verbs was correlated highly with the frequency of the word in our training corpus (nouns, $r = .7353$, verbs, $r = .6828$). Word Prediction accuracy was also highly correlated with the frequency of the target word ($r = .6266$). This is not surprising in either case, since the ability of the network to learn a pattern is dependent upon how often it sees it.

Discussion

We expected that this experiment would demonstrate the advantage of including meaningful features in the word learning and word prediction process, and this is exactly what we found. While the absolute prediction accuracy of the networks is not yet very good, there is a small but significant difference in prediction accuracy

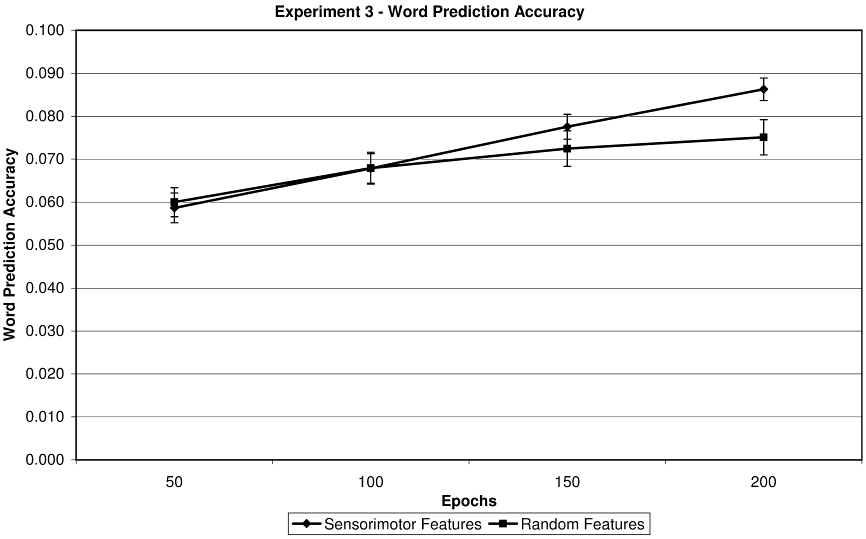


Fig. 4. Mean word prediction performance for Experiment 3. The number of networks in each condition is 10. Error bars indicate standard error.

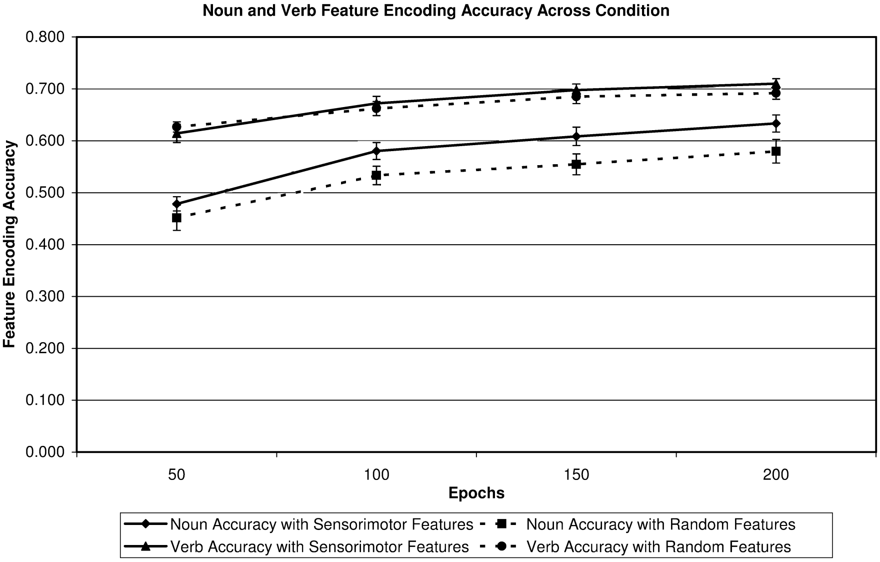


Fig. 5. Noun and Verb feature encoding accuracy from Experiment 3. These two output layers were performing a mapping from the phonetic features of a word to the semantic features of a word. The number of networks in each condition is 10. Error bars indicate standard error.

Table 3
Output accuracy from sample network at 500 epochs, during training

	Noun features encoding	Verb features encoding	Word prediction
Accuracy	65.535%	75.251%	28.030%
Number of items	60 grounded nouns in this corpus	49 grounded verbs in this corpus	529 words in this corpus total

between the two conditions at the completion of training.

Using the exact match error criterion, a difference of 13% in word prediction accuracy (a simple measure of grammar learning abilities) is evident at the final point of training, and the difference between the two conditions' average accuracy curves is increasing over the latter portion of training, as demonstrated by the repeated-measures ANOVA.

The categorical match error criterion produces a similar result (8.2% difference between the two conditions) at the final epoch of training, and is also significant. However, given that using the exact match measure is much easier to calculate than the categorical match measure, and does not involve issues such as the choice of the right level of grammatical categories to use, etc., it seems appropriate that we have been using the more conservative exact match grammatical accuracy measure. Still it is interesting to see that the results do not depend on the choice of grammatical accuracy criterion.

Also, the ability of the network to map the phonetically presented word inputs to semantic features (an analogue of lexical learning) is significantly different between the two conditions, at least for nouns. The fact that this effect was not significant for verbs may be due to the fact that fewer of them were grounded in our training corpus (60 nouns versus 49 verbs) and the fact that the network has more exposure to nouns (since most simple sentences contain only one verb, but several nouns). Similarly, the fact that overall, the accuracy for verbs is better than nouns (a counterintuitive result) is likely related to the relative sparseness of the multidimensional space in which the verbs are represented by their features. When the network output is forced via the Euclidian-distance decision rule to select one verb form as its match, there are fewer neighbors in the verb space, and greater base likelihood of selecting the correct one than in the noun feature space.

Overall, these results demonstrate the ability of sensorimotor features to improve both "lexical learning," the process of mapping word forms to conceptual representations, and a simple aspect of "grammatical learning," the process of sequence learning.

Of course, it was also important to simply show that the network was able to perform the task of producing sensorimotor features at output that correspond to the meaning of the word presented phonetically at input. This finding will be the basis for further studies examining the potential of this network to exhibit a "propagation of grounding" effect—the ability of the network to learn to produce meaningful features for novel, ungrounded word forms.

General discussion and conclusions

The preceding experiments (Experiments 1 and 2) have demonstrated that representing concepts in terms of sensorimotor features captures important aspects of the semantic meaning of those concepts, and that this knowledge is structured in meaningful ways (although this is clearer for nouns than for verbs). We have also shown (Experiment 3) that the inclusion of these sensorimotor features as semantic representations of words in a model of language acquisition can improve performance on both lexical learning (11% difference for nouns) and grammatical learning (13% difference). These results demonstrate that having sensory and motor knowledge of objects and events in the environment is a significant advantage when trying to acquire language for the first time, for networks and presumably for children.

We can characterize these results partly in terms of the artificial language learning literature, as well as in relation to results from other language acquisition models investigating other language cues (e.g., prosody Christiansen, Allen, & Seidenberg, 1998). In performing its word prediction task (our 'syntax' task) our network essentially has two sources of information upon which to operate, word form (represented phonologically) and sensorimotor semantics. Research on the effects of multiple cues during artificial language learning indicates that having at least one other cue in addition to the transitional probabilities of word symbols increases the learnability of grammatical classes whether this cue is linguistic or extralinguistic (e.g., McDonald & Plauche, 1995).

In our network, however (and contrary to the random letter strings often used in artificial language learning experiments), the word representation is more than just an arbitrary symbol, it is a full phonological representation. Given that aspects of phonology can serve as linguistic markers, it was possible that our word forms served both as the symbols to be sequenced and a linguistic cue to their grammatical usage. Thankfully, the random condition in Experiment 3 controls for this possible cue (as it is the same in both experimental and random conditions) leaving us confident that the effect we found is in fact due to the inclusion of sensorimotor semantics (an arguably extra-linguistic cue in this case). Of course, the degree of transparency, or salience, of the multiple cues involved is also relevant (McDonald & Plauche, 1995). In our experiment, the transparency of the semantic representations is very high, as they are an explicit target of network operation. Phonological markers, if any, are much less transparent due to their relatively hidden status within the word forms. Normally, almost all learning in such multi-cue situations of differential transparency is directed towards the highly transparent cue (McDonald & Plauche, 1995), which again leaves us confident that the difference we detected was due to the sensorimotor semantic representations.

This is not to say that we could not to extend our model to incorporate other learning cues, profiled in such a way as to be easily combined with semantic cues. Indeed, this notion of multiple interacting cues is exactly what researchers like Seidenberg and MacDonald (2001) advocate. One likely cue that would be a good candidate for inclusion in our model would be prosody or syllabic stress (e.g., Christiansen et al., 1998)

Another important caveat is that even though Experiment 3 used a corpus which was a concatenation of transcribed mother-to-child speech taken from the CHILDES database (Bates et al., 1988; Carlson-Luden, 1979; MacWhinney, 2000), only 60 nouns and 49 verbs were actually represented in our vocabulary of 352 grounded early nouns and 90 grounded early verbs. Why do so few of the corpora's 529 words overlap with our grounded words? One possible reason is related to the situation in which the speech was originally elicited; a relatively constrained joint-play situation rather than natural in-home childhood activities, where more of the words from our MCDI set would presumably be encountered and discussed. Even so, the fact that so few words were grounded in sensorimotor features is not a problem for our account. Consider our semantic representations as a cue to grammatical learning again, as discussed above. Had more of the corpora's vocabulary of 529 words been grounded, the usefulness of the semantic cue to the network in performing word prediction would have been more obvious. As it was, most words had no semantic targets, making it harder for the network to learn this cue. This is undoubtedly part of the reason for the small size of the effect we found. Had more words been grounded, the effects of including the sensorimotor features would likely have been much larger and more obvious.

These experiments were performed using the most naturalistic corpus we could find, which was intended to allow for good input representativeness (Christiansen & Chater, 2001) on the part of the model, the idea being that it is receiving the same sort of input that the child might, and so the model's results would be extendable more readily to the case of child language acquisition. This goal certainly still holds; we *can* be confident that this effect of pre-linguistic conceptual knowledge on grammatical learning should appear in children as well as the network. At an extreme level, it is obvious that it has to. If a child has no meaning representations for any of the words that he or she is hearing in speech, then the grammar of the language will be impossible to learn. This is McClelland's "learning a language by listening to the radio" criticism (Elman, 1990). Thus, the more words whose meaning is known that occur in the speech stream, the more the grammar is inferable, and the more easily that novel words can be understood. Experimental evidence of this process has been discussed earlier (Gillette et al., 1999).

However, the naturalistic corpus that we used for these simulations was *not* a very grammatical corpus! Upon examination, the mother-to-child speech contains very few proper sentences, and very many partial sentence fragments, repetitions of words, attention-eliciting verbal behaviors (e.g., "look at this, what is this," etc.), and as mentioned previously, only a minority of the words were grounded in sensorimotor features (as discussed above, this may be related to the nature of the original somewhat constrained joint-play scenario from which the speech was transcribed, and thus we make no claims regarding the overall implications of 'motherese' for grammar learning). Thus, while the use of this 'naturalistic' corpus may make us confident in generalizing from the network's behavior to that of children, it makes it very difficult to produce that network behavior in the first place, hence the low overall accuracy found in Experiment 3. In fact, by including only mother-to-child speech, this training corpus is much more impoverished than what children would be exposed to, since children also overhear more grammatical adult-to-adult speech. We would thus expect larger grammatical differences between experimental and control conditions to be evident with training corpora that were more grammatical.

Of course, the other interpretation is that the lack of overlap between the 529 words of our mother to child corpus and the 442 words of our sensorimotor semantic representations actually calls our effect into question. One might argue that this lack of overlap shows that children actually do not receive enough grounded input to make it a viable cue for language learning. There are several indications that this is not the case, however. One is that in pilot work, we observed that the more words in the training corpus that were grounded, the larger was the effect of sensorimotor features on language learning. This extends to the point of arguably easiest word learning, when all other words in a situation are known and grounded and only one is not. As discussed before, this is essentially the "fast-mapping" paradigm of word learning (Bloom, 2000), although we are not claiming that networks can do "fast-mapping" per se, merely that the behavior that we see in the network is a less extreme form of the phenomenon, essentially just a degree of facilitation of learning. Furthermore, in preliminary work on the propagation of grounding effect using simpler semantic representations, we found that the larger the percentage of the words in the corpus that were grounded, the more likely a novel word was to acquire an appropriate meaning. As discussed below, this experiment has not yet been conducted on these particular rich semantic representations, however. Further, as discussed above, the MCDI words are primarily words that children would encounter in normal daily home life, and should be more fully represented in speech taken from such a context. In any event, we believe that the combination of these factors serves to support our interpretation of our results.

Sensorimotor features can also be used in other ways in models of language. They might be particularly useful in modelling in detail the process of word learning. If as Bloom (2000) suggests, children learn the meanings of words through attention to what the caregiver is attending to, then combining feature representations with phoneme-by-phoneme speech representations might be a network analogy. This would help the network to learn to bind individual phonemes into words, using the constancy of sensorimotor features (as an analogue to focused joint attention with a caregiver) to determine that all these phonemes apply to the same perceived object. In unpublished work, we have begun to examine exactly this.

Another advantage of sensorimotor features relates to word sense disambiguation: this kind of meaning representation may be used to disambiguate multiple senses of a word encountered in text, through the operation of feature prediction in concert with word prediction. That is, if the network is predicting the word "bank" next, by examining the features it is predicting at the same time we might be able to tell whether it means to output "a place to store money" or "the edge of a river". In fact, it has been suggested (Ken McRae, Private Communication, 2004) that this is arguably the most interesting usage of these features.

Now that we have a set of explicitly grounded sensorimotor features for the earliest words, a question naturally arises: do we need to derive featural ratings for every concept that the network is exposed to? Fortunately, that should not be necessary. As discussed previously, evidence indicates that only the child's earliest words are fully grounded in sensory experience (Gillette et al., 1999); in fact it is the early words' very imageability and accessibility to observation that leads them to *be* the first words generally learned by children. As lexical learning progresses, less and less imageable (i.e. more abstract) words are experienced and learned. Also, the learner is exposed to novel words in speech or text that are not directly grounded in immediate sensory experience. Both of these sorts of words can be grounded only indirectly by association with other more imageable words in the context. Therefore, if we empirically generate the sensorimotor features for the most imageable, earliest words in children's lexicons, we can reasonably expect that later words will be effectively grounded via their relationships to these earlier words. In the neural network model, novel words presented to the model without accompanying sensory input should begin to elicit the appropriate sensorimotor features due to similarities to other concepts or words that share context or usage (see Howell et al., 2001; for a detailed discussion). This is our "propagation of grounding" process. While the present work does not directly investigate this process, the above demonstration of the contribution of sensorimotor features to lexical and grammatical learning was a necessary first step. We are now experimenting with networks designed to investigate this propagation of grounding more directly.

Acknowledgments

This research was supported by a National Science and Engineering Research Council (NSERC) postgraduate scholarship and an Ontario Graduate Scholarship to SRH, an NSERC postgraduate scholarship to DJ, and a research grant from NSERC to SB. SRH would like to thank George Lakoff for inspiring the use of sensorimotor features, and Lee Brooks for helpful advice throughout this work. We would also like to thank three anonymous reviewers for their valuable assistance with the manuscript.

Appendix A. Instructions and dimensions for Experiment 1 (Excerpts from subject instructions)

On the following pages are a series of various concepts or words, such as "dog," or "kettle.' For each of the concepts/words, there is a list of *features.* Please rate *each* concept on *each* feature on a scale of 0 to 10. Try to picture the object or concept mentally as you are making your rating, including its sounds, smells, motions, etc. . . . you should try to limit yourself to the knowledge of the world that an average pre-school child would have. For example, for size, do not compare the concept in question to a microscopic bacteria or to a mountain. You might limit your comparison group to anywhere from the size of a pea (tiny = 0) on up to the size of a house (extremely large = 10), for example. . .

Noun semantic dimensions or features

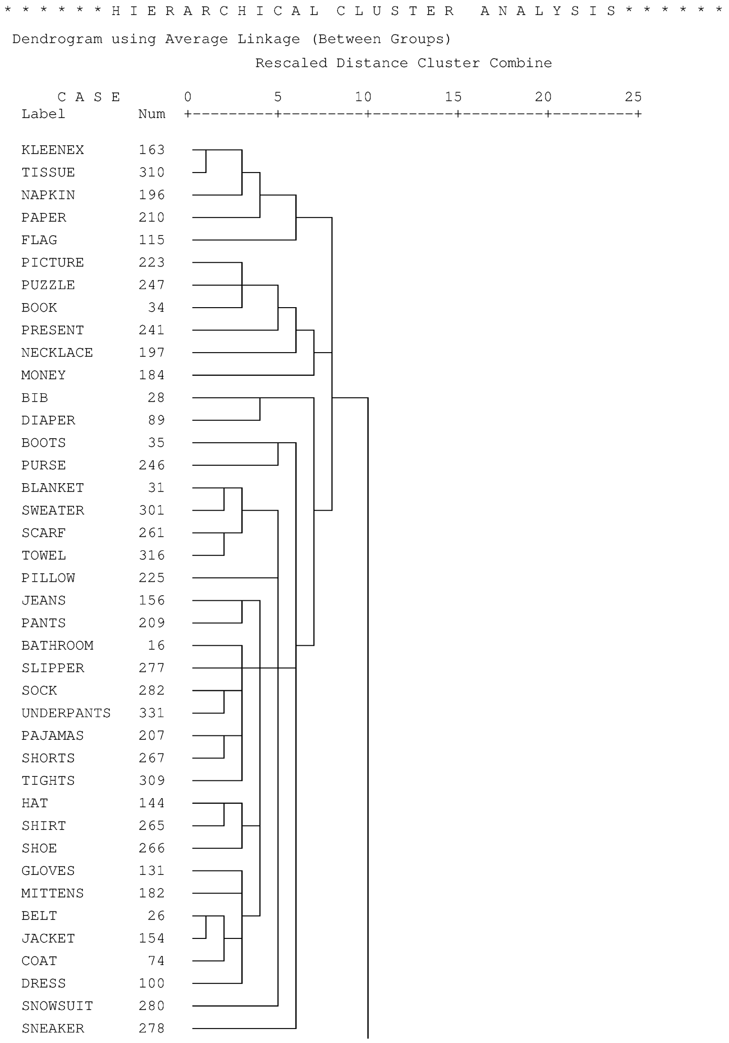
Size
Weight
Strength
Speed
Temperature
Cleanliness
Tidiness
Brightness
Noise
Intelligence
Goodness
Beauty
Width
Hardness
Roughness
Height
Length
Scariness[*]
Colourfulness
is_black
is_blue
is_brown
is_gold
is_green
is_grey
is_orange
is_pink
is_purple
is_red
is_silver
is_white
is_yellow
is_conical
is_crooked
is_curved
is_cylindrical
is_flat
is_liquid
is_rectangular
is_round
is_solid
is_square
is_straight
is_triangular
has_feathers
has_scales
has_fur
is_prickly
is_sharp
is_breakable
made_of_china
made_of_cloth
made_of_leather
made_of_metal
made_of_plastic
made_of_stone
made_of_wood
climbs
crawls
flies
leaps
runs
swims
breathes
drinks
eats
makes_animal_noise
sings
talks
has_4_legs
has_a_beak
has_a_door
has_a_shell
has_eyes
has_face
has_fins
has_handle
has_leaves
has_legs
has_paws
has_tail
has_teeth
has_wheels
has_whiskers
has_wings
is_annoying
is_comfortable
is_fun
is_musical
is_scary [*]
is_strong_smelling
is_young
is_old
is_comforting
is_lovable
is_edible
is_delicious

[*] Note that due to an oversight, is_scary and scariness are both included in this set. We do not expect that this has any effect on the results.

Instructions for Part 1. Please enter a value between 0 and 10, with 5 being in the middle of the two opposites (first 19 dimensions, Size—Colourfulness)

Instructions for Part 2. Please enter a value between 0 and 10. A value of 10 means the feature is ALWAYS true or ALWAYS present, a value of 0 means that it is NEVER true or present, and a value of 5 means that it is true about 50% of the time, or for 50% of the instances of that concept. Example: if you think that 60% of the time an apple is red, then rate apples a 6 on is_Red.

Appendix B. Sample of noun cluster analysis



```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

 Dendrogram using Average Linkage (Between Groups)
                           Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label       Num  +---------+---------+---------+---------+---------+

  KLEENEX      163
  TISSUE       310
  NAPKIN       196
  PAPER        210
  FLAG         115
  PICTURE      223
  PUZZLE       247
  BOOK          34
  PRESENT      241
  NECKLACE     197
  MONEY        184
  BIB           28
  DIAPER        89
  BOOTS         35
  PURSE        246
  BLANKET       31
  SWEATER      301
  SCARF        261
  TOWEL        316
  PILLOW       225
  JEANS        156
  PANTS        209
  BATHROOM      16
  SLIPPER      277
  SOCK         282
  UNDERPANTS   331
  PAJAMAS      207
  SHORTS       267
  TIGHTS       309
  HAT          144
  SHIRT        265
  SHOE         266
  GLOVES       131
  MITTENS      182
  BELT          26
  JACKET       154
  COAT          74
  DRESS        100
  SNOWSUIT     280
  SNEAKER      278
```

(Rest of graph omitted due to lack of space)

Appendix C. Forms and instructions for Experiment 2

C.1. Pilot study—verb features generation

Please list as many features/aspects of the following verbs as you can. There is no rush, please take the time to think about and visualize (if possible) the word in question. Try to focus on physically observable aspects of that verb, rather than on other words, nouns, etc, that it tends to occur with. A "feature" does not have to be a single word, so for "fly" the features might be something like:

Requires wings
Goes fast

Travels from point a to point b
Moves through the air
Etc…

(Verbs listed for rating included prototypical ‘light’ verbs such as go, put, move, hit, etc.)

C.2. Experiment 2—Excerpts from subject instructions

On the following pages are a series of various actions or verbs, such as “**hit**,” or “**run**.” For each of the actions/verbs, there is a list of *features*. Please rate *each* verb on *each* feature on a scale of 0 to 10. Try to picture the object or concept mentally as you are making your rating… In general for these ratings, you should limit yourself to the experience that a pre-school child might have, that is, very basic physical understandings of themselves and their actions.…

Verb semantic dimensions or features

Joint motion

Toes
ankles
knees
hips
torso
shoulders
elbow
wrist
fingers
neck
head
face
eyes
eyebrows
nose
mouth
lips
tongue
requires a specific overall bodily position?
degree of overall body contact involved
horizontal motion involved
vertical motion involved
optimum size of actor

Sensory perceptions/physical observations

noisiness (0 = silence)
perception—Auditory
perception—mental
perception—Smell
perception—Taste
perception—Touch
perception—Visual
speed (10 = fastest)
suddenness (0 = totally expected)
tightness (0 = no hold)
agitation (physical)
balance (0 = totally unsteady)

Physical states

decreases agitation
decreases energy
decreases hunger
decreases thirst
decreases tiredness
increases agitation
increases energy
increases hunger
increases thirst
increases tiredness
reactiveness (0 = unreactive to stimuli)
tension (0 = completely relaxed)

Mental state features

aggression (0 = complete passivity)
attention (0 = oblivious to this stimulus)
awareness (0 = unaware of anything)
control (0 = completely accidental)
pleasurable (0 = not at all)
painful (0 = not at all)
purposeful (0 = completely unintentional)

Temporal features

starts something else
ends something else
duration of action (0 = instantaneous)
periodic action (0 = single action)
time pressure involved (e.g., verb “race”)

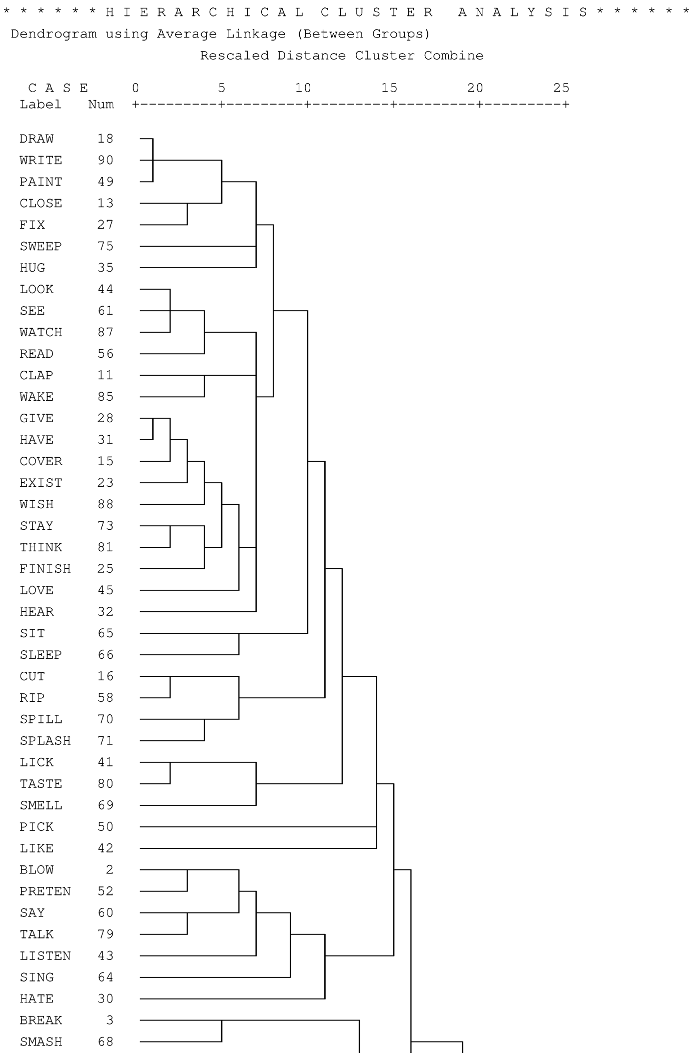
Physical requirements/characteristics

amount of contact involved between actor and object
involves container/containing
involves supporting something
forcefulness
requires physical object
requires a surface
strength involved
involves a trajectory from source to goal

Appendix C (*continued*)

Physical effects
interrupts a path or trajectory
causes damage
distance typically
conjoins things
divides things
consumes (e.g., uses up like in "burn")
creates (something new)
destroys
displaces other object (and takes its place)
creates disorder/untidiness
creates order/tidiness
closes/closes down
opens/opens up
change is involved (0 = totally static)
transference of something
assembles things
disassembles things

Appendix D. Subset of verb cluster analysis



```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *
 Dendrogram using Average Linkage (Between Groups)
                      Rescaled Distance Cluster Combine

   C A S E     0         5        10        15        20        25
  Label   Num  +---------+---------+---------+---------+---------+

  DRAW     18
  WRITE    90
  PAINT    49
  CLOSE    13
  FIX      27
  SWEEP    75
  HUG      35
  LOOK     44
  SEE      61
  WATCH    87
  READ     56
  CLAP     11
  WAKE     85
  GIVE     28
  HAVE     31
  COVER    15
  EXIST    23
  WISH     88
  STAY     73
  THINK    81
  FINISH   25
  LOVE     45
  HEAR     32
  SIT      65
  SLEEP    66
  CUT      16
  RIP      58
  SPILL    70
  SPLASH   71
  LICK     41
  TASTE    80
  SMELL    69
  PICK     50
  LIKE     42
  BLOW      2
  PRETEN   52
  SAY      60
  TALK     79
  LISTEN   43
  SING     64
  HATE     30
  BREAK     3
  SMASH    68
```

(Rest of graph omitted due to lack of space)

References

Bailey, D., Feldman, J., Narayanan, S., & Lakoff, G. (1997). Modelling embodied lexical development. In *Proceedings of the cognitive science society*, 1997.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge, MA: Cambridge University Press.

Bates, E., & Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language*. New Jersey: Lawrence Erlbaum Associates.

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

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In Dietrich & Markham (Eds.), *Cognitive dynamics: Conceptual change in humans and machines*.

Carlson-Luden, V. (1979). *Causal understanding in the 10-month-old*. Unpublished doctoral dissertation. University of Colorado at Boulder.

Christiansen, M. H., & Chater, N. (2001). *Connectionist Psycholinguistics*. Westport, Ct.: Ablex Publishing.

Christiansen, M., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics, 21*, 95–115.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development, Vol. 2: Language, thought, and culture* (pp. 301–334). Hillsdale, NJ: Lawrence Erlbaum.

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

Glenberg, A. M., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*, 558–565.

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

Goldberg, A. (1999). The emergence of argument structure semantics. In B. MacWhinney (Ed.), *The emergence of language*. New Jersey: Lawrence Erlbaum Associates.

Hinton, G. E., & Shallice, T. (1991). Lesioning a connectionist network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74–75.

Howell, S. R., & Becker, S. (2000). Modelling language acquisition at multiple temporal scales. In *Proceedings of the 22nd annual conference of the cognitive science society* (p. 1031). 2000.

Howell, S. R., & Becker, S. (2001). Modelling language acquisition: Grammar from the lexicon? In *Proceedings of the 23rd annual conference of the cognitive science society conference* (pp. 429–434). 2001.

Howell, S. R., & Becker, S. (2005). SRNEngine: A Windows-based neural network simulation tool for the non-programmer. *Manuscript in Preparation.*

Howell, S. R., Becker, S., & Jankowicz, D. (2001). Modelling language acquisition: Lexical grounding through perceptual features, In *Proceedings of the 2001 developmental and embodied cognition conference*, July 31, 2001.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer-Verlang.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago and London: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–242.

Landauer, G. T., Laham D., & Foltz, P. (1998). Learning Human-like knowledge by singular value decomposition: A progress report.

Langer, J. (2001). The mosaic evolution of cognitive and linguistic ontogeny. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge: Cambridge University Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago: University of Chicago Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology, 20*, 121–157.

McDonald, J. L., & Plauche, M. (1995). Single and correlated cues in an artificial language learning paradigm. *Language and Speech, 38*(3), 223–236.

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*, 99–130.

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

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review, 99*, 587–604.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11–28.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (pp. 318–362). Cambridge, MA: MIT press.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences, 3*, 417–424.

Seidenberg, M. S., & MacDonald, M. C. (2001). Constraint satisfaction in language acquisition and processing. In M. H. Christiansen & N. Chater (Eds.), *Connectionist psycholinguistics* (pp. 177–211). Westport, CT: Ablex Publishing.

Smith, L. B. (1999). Children's noun learning: How general learning processes make specialized learning mechanisms. In B. MacWhinney (Ed.), *The emergence of language*. New Jersey: Lawrence Erlbaum Associates.

Smith, L. B., & Jones, S. S. (1993). The place of perception in children's concepts. *Cognitive Development, 8*(2), 113–139.

Vinson, D., & Vigliocco, G. (2002). A semantic analysis of grammatic class impairments: Semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics, 15*(3–5), 317–351.