Connection Science
Vol. 17, No. 3–4, September–December 2005, 381–397

Taylor & Francis
Taylor & Francis Group

The emergence of links between lexical acquisition and object categorization: a computational study

CHEN YU*

Department of Psychology and Cognitive Science Program, Indiana University, Bloomington, IN 47405, USA

Language is about symbols, and those symbols must be grounded in the physical world. Children learn to associate language with sensorimotor experiences during their development. In light of this, we first provide a computational account of how words are mapped to their perceptually grounded meanings. Moreover, the main part of this work proposes and implements a computational model of how word learning influences the formation of object categories to which those words refer. This model simulates the bi-directional relationship between word and object category learning: (1) object categorization provides mental representations of meanings that are mapped to words to form lexical items; (2) linguistic labels help object categorization by providing additional teaching signals; and (3) these two learning processes interplay with each other and form a developmental feedback loop. Compared with the method that performs these two tasks separately, our model shows promising improvements in both word-to-world mapping and perceptual categorization, suggesting a unified view of lexical and category learning in an integrative framework. Most importantly, this work provides a cognitively plausible explanation of the mechanistic nature of early word learning and object learning from co-occurring multisensory data.

Keywords: Word learning; Object categorization; Computational modelling

1. Introduction

Anyone who has watched an infant grow into an adult can attest to the changing nature of cognitive functions during human development. Infants are born with limited perceptual, linguistic and cognitive abilities. They rapidly develop skills and acquire knowledge based on their sensorimotor experiences with the physical world. One of the most fundamental capacities is to talk about what they see. Starting from scratch, infants gradually acquire a vocabulary and grammar. Although this process continues throughout childhood, the crucial steps occur early in development. By the age of 3 years, most children have incorporated the rudiments of grammar and are rapidly increasing their vocabulary. Exactly how they accomplish this is uncertain, but it is clear that children make use of information both in the language around them (Saffran *et al.* 1996) and in the extralinguistic world. Although the unprocessed audio stream is rife with ambiguity, recent tests have shown that babies are sensitive to correlations

*Email: chenyu@indiana.edu

Connection Science
ISSN 0954-0091 print/ISSN 1360-0494 online © 2005 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09540090500281554

in sounds (Saffran *et al.* 1996). Children also use prodigious amounts of information about the world in the language process. This extralinguistic information is needed because language is about symbols and the meanings of those symbols must be grounded in the physical world. Children develop based on their sensorimotor experiences with the physical environment. Different levels of abstraction are necessary to encode efficiently those experiences, and one vital role of the human brain is to bridge the gap from embodied experience to its expression as words in a language. This was termed symbol grounding by Harnad (1990). In this paper, we focus on grounding object names (a specific set of words) in visual perception. Object names are focused on because early vocabularies are largely made up of names applied to solid objects (Smith *et al.* 2002). Object perception is focused because objects are primary perceptual and cognitive units during infancy (Spelke *et al.* 1995). For the developing child, solving the problem of grounding object names requires advances in both word learning (what an object name refers to) and object categorization (how to group instances of objects into categories). Importantly, progress in object categorization changes rapidly at the same time as children's object name vocabularies also expand rapidly (Smith 2003), suggesting possible links between these two processes.

1.1 *Prelinguistic object categorization*

To learn object names, infants must build object categories. This provides the starting point for learning first words. Recent research suggests that even prelinguistic infants have complex categorization abilities. Early in development, infants exhibit the ability to form categories by treating discriminable individuals as members of an equivalence class (Mareschal and Quinn 2001). This task is most often termed *perceptual categorization*. Most studies report evidence of infant categorization by 7–10 months, but Quinn *et al.* (1993) reported that infants as young as 3 months can form basic-level object categories. Moreover, Landau *et al.* (1998) suggested that infants respond to categories on the basis of visual features in the exemplar and shape similarity plays an important role in early object categorization. However, most object categories contain sets of objects that need not share visual similarity. Therefore, an intriguing question is how might children learn to form these sorts of conceptual categories and then map object names to cover such sets at the later stage of development?

1.2 *Early word learning*

Infants begin to comprehend words at 6–10 months of age. The largest proportion of most children's first 50 words consists of names for objects, such as food, clothing and toys. Gillette *et al.* (1999) provided strong evidence that learnability of a word is based primarily upon its imageability or concreteness. For instance, the infant may learn that the sound 'car' is associated with visual features (colour, shape, texture, etc.) of a toy car because her mother utters the sound 'car' when she is looking toward it. Smith (2000) proposed that word learning is initially a process in which children's attention is captured by objects or actions that are the most salient in their environment. Those objects or actions are then associated with acoustic patterns voiced by an adult. In everyday situations, the infant perceives connected spoken utterances instead of isolated spoken words. A spoken utterance usually consists of several words, each of which may have a referent in the physical environment. An example illustrating natural interactions between a mother and her infant is as follows: the mother holds a cow toy in her left hand and a sheep toy in her right hand. She produces a spoken utterance 'here is a cow and here is a sheep' while the infant is looking toward her hands. In this example, all the spoken words, such as 'here', 'is', 'a', 'cow', 'sheep' and 'and', are co-occurring

with the visual objects 'cow' and 'sheep'. For the purpose of word learning, the infant needs to find which items on the language side are likely to associate with which items obtained from visual perception. Thus, a central issue is the *word-to-world mapping*—how to discover correct word–meaning pairs from multiple co-occurrences between words and things in the environment.

1.3 *Links between object naming and categorization*

Perceptual similarity plays a critical role in early language learning by grouping objects into categories that will support early object naming. It has been shown that children generalize object names to new objects often on the basis of similarity in shape (Clark 1973, Landau *et al.* 1997) and in texture (Smith and Heise 2000). Behl-Chadha (1996) (see also Quinn *et al.* 2001) demonstrated that even 4-month-old infants can distinguish objects in basic-level categories based on shape, suggesting that knowledge of some object categories is constructed before object name learning begins.

Perceptual similarity alone seems to be insufficient to explain object naming despite its strong role in early object categorization. An object name is used to refer to many perceptually different entities. For instance, the same linguistic label *dog* can refer to different kinds of dogs with different colours, shapes and textures. Moreover, young learners are able to extend that linguistic label to individual objects of the same kind that they have never seen. Thus, they go beyond word-to-referent and word-to-object mappings toward building word-to-category mappings.

Landau *et al.* (1998) (see also Smith *et al.* 1996) showed that young children have a tendency to generalize novel object names on the basis of an object's shape. This shape bias is found only in naming contexts, not in the conditions where the objects are not named. Thus, it seems that salient perceptual similarity of objects can direct children's generalization of category labels on novel objects (Gentner 1978, Smith *et al.* 1996, Landau *et al.* 1998). In this way, young children learn to associate linguistic labels with concepts—mental representations of meanings, not just concrete instantiations of these concepts. A recent study by Yoshida and Smith (2005) also showed that the presence of correlated linguistic cues enhances children's learning about the perceptual cues to object categorization. Furthermore, Sloutsky and Lo (1999) showed that, if two entities share the same label, young children are more likely to conclude that these entities look alike. Xu (2002) showed that 9-month-old infants could use two distinct linguistic labels to facilitate object individuation. In Waxman and Markow (1995), the experimenter offered four different toys from a given category to infants in the familiarization phase and manipulated the audio stimuli to include a no-word condition, a noun condition and an adjective condition. Infants in both the noun and adjective conditions formed object categories, while infants in the no-word condition failed to detect the categories. This study offered clear evidence that linguistic labels embedded in continuous speech support the establishment of object categories.

Thus, there are data showing the independent contribution of similarity-based categorization to naming and of naming to categorization. However, these two factors are also likely to interact. This leads us to ask the following questions. (1) How does language change the process of categorizing objects? (2) How does object categorization, in turn, drive the establishment of word-to-category mapping? As shown above, many researchers agree that language can shape categories, and there is a variety of empirical studies to support the idea. However, these experimental studies do not offer an explanation of the interaction between object naming and object categorization in terms of internal mechanisms. They cannot answer the question of *how* these two learning processes interplay with each other. In this paper, a computational

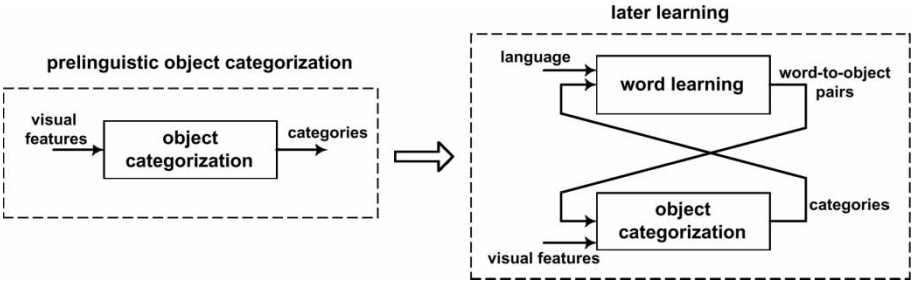

Figure 1. Object categorization before language acquisition is based purely on perceptual similarity. With the introduction of linguistic information, word learning and object categorization interact with each other and form a developmental feedback loop.

mechanism is proposed that attempts to study the influence of object categorization on word learning as well as the influence of word-to-object mappings on object categorization. In our proposal, as shown in figure 1, perceptual similarity plays a key role in the early development of object categorization and serves as an initial basis for early acquisition of object names. At a later stage, with more word-to-object pairs accumulated, children start to use linguistic labels as additional teaching signals to guide object categorization. In terms of learning mechanisms, children start with unsupervised learning based on perceptual similarity and gradually transform to semi-supervised learning based on correlated linguistic labels as additional teaching signals. This paper renders the idea as a working model.

In summary, a young child seems effortlessly to perceive, categorize and name objects. Despite the apparent ease with which he accomplishes this, the mechanisms supporting these capacities involve perceptual, linguistic and cognitive systems, and are likely to be complex. Recently, a new trend in cognitive development is to use computational modelling as a complementary methodology in the study of the mind (Plunkett *et al.* 1992, Elman *et al.* 1996, Cangelosi and Parisi 1998, Brent 1999, Regier 2003). Potential models can be implemented as computer programs. The strengths of such models are in providing explicit mechanisms of cognition as well as generating detailed empirical predictions. In light of this, a computational model is developed here to explore the links between word and object category learning. More specifically, the model provides a formal account of: (1) how object names are grounded in visual perception; (2) how word-to-world mappings facilitate object categorization; and (3) how these two advances bootstrap each other and form a developmental feedback loop. In the model, object categorization is based not only on perceptual similarities (appearance features), but also on co-occurring linguistic labels used as potential teaching signals to guide the categorization. To our knowledge, this model is the first that tackles the word-to-world mapping and perceptual categorization problems within a single framework while processing sensory data collected from a natural environment. The specific contribution is to show that, based on an integrative view, these two learning problems can be fundamentally simplified by exploring the possibility of utilizing the spatio-temporal and cross-modal constraints in multimodal data.

2. Related work

This section provides a brief review of related work on both modelling language acquisition and multimodal learning.

2.1 *Models of word learning*

Computational investigations of language acquisition have recently received considerable attention. MacWhinney (1989) applied competition theory in building an associative network that was configured to learn which word among all possible candidates refers to a particular object. Siskind (1996) developed a mathematical model based on cross-situational learning and the principle of contrast, which learns word–meaning associations when presented with paired sequences of pre-segmented tokens and semantic representations. Regier's (1996) work focused on grounding lexical items that describe spatial relations in visual perception. Tenenbaum and Xu (2000) developed a computational model based on Bayesian inference that can infer meanings from one or a few examples without encoding the constraint of mutual exclusion.

The importance of embodiment in word acquisition is featured in a variety of sources (Plunkett *et al.* 1992, Steels 1997, Cohen *et al.* 2001, Roy and Pentland 2002, Yu *et al.* 2003, Weng *et al.* 2003) (also see Lungarella *et al.* (2004) for a good survey of developmental models). Among them, Plunkett *et al.* (1992) built a connectionist model of word learning in which a process termed auto-association maps preprocessed images with linguistic labels. The linguistic behaviour of the network exhibited non-linear vocabulary growth (vocabulary spurt) that is similar to the pattern observed in young children. Steels and Vogt (1997) reported experiments in which autonomous visually grounded agents bootstrap meanings and language through adaptive language games. They argued that language is an autonomous evolving adaptive system maintained by a group of distributed agents without central control and a lexicon may adapt to cope with new meanings that arise. Cangelosi *et al.* (2000) showed that, in a simulated artificial environment, a better categorization could be achieved by utilizing associated language-like symbols compared with the approach based purely on sensorimotor signals. Roy and Pentland (2002) used the correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. The learning algorithm is based on cross-modal mutual information to discover words and their visual associations. In Yu *et al.* (2003) and Yu and Ballard (2004), egocentric multisensory data were used first to spot words from continuous speech and then to associate action verbs and object names with their perceptually grounded meanings. The central idea is to utilize body movements as deictic references to associate temporally co-occurring data from different modalities. These models focus on learning to associate words with meanings. Few of them (but also see Cangelosi *et al.* 2000) attempt to simulate the interdependence between object name learning and object categorization.

2.2 *Multimodal learning*

Biological studies have shown that sensory information is shared not just in a final integrative stage, but across all levels of perceptual processing. A convincing demonstration of audio-visual correlations that radically change an auditory perception is the *McGurk effect* (McGurk and MacDonald 1976). McGurk and MacDonald showed that paired auditory and visual stimuli can interact and produce a unified result different from either the actual auditory input or the visual sensory input. In light of this, several computational studies have considered using the relationships between inputs of different modalities to facilitate the learning problem. Among them, the works in de Sa and Ballard (1998) and Becker (1996) are most instructive because both works provided a means of using the information in one channel to modulate learning in another channel, rather than simply merging the outputs of multiple streams. deSa and Ballard (1998) related this idea mathematically to the optimal goal of minimizing the number

of misclassifications in each modality and applied it to derive an algorithm for two piecewise linear classifiers, in which one uses the output of the other as supervisory signals. The method in Becker (1996) was derived from information theoretic principles and is based on the idea of maximizing the mutual information between the outputs of different neural network modules. The approach maximizes some measure of agreement between the outputs of two groups of units that receive inputs physically separated in space, time or modality. However, both methods assume that the data from different modalities are perfectly synchronized (e.g. lip motion and speech). Clearly, this kind of synchrony does not exist in the context of word learning (Gleitman 1990).

3. The model

In this section, we first introduce the general principle of our probabilistic model using a simple example. A detailed description is provided in the following subsections.

3.1 *General description*

Our method is based on spatio-temporal and cross-modal regularities between words and visual objects, which consists of three steps, as illustrated in figure 2.

- Clustering: visual features extracted from the appearances of objects are quantified to form prototypes based on their perceptual similarity in a feature space.
- Categorization: visual prototypes are grouped into object categories.
- Association: the association probabilities between words and object categories are estimated.

Now we have a difficult chicken-and-egg problem. Without the results of categorization, we cannot estimate association probabilities between words and meanings (object categories). Likewise, without the results of association, we cannot find and then use corresponding linguistic labels as teaching signals to guide the categorization of visual prototypes into object categories. Our solution is first to cluster visual objects into prototypes based on perceptual features, as described in section 3.2. In the next step, each visual prototype is assigned to a distinct object category. In this way, we can estimate the initial association probabilities between words and object categories, as presented in section 3.3. Based on the results of initial training, the third step iteratively merges visual prototypes into a new set of object categories so that one category includes multiple visual prototypes. More specifically, both perceptual similarity and associated linguistic labels are used in categorizing visual prototypes; and new categorization leads to a re-estimate of association probabilities between words and a new set of object categories. In this way, perceptual categorization and word-to-world mapping processes bootstrap each other and form a developmental feedback loop. In addition, we perform a similar learning procedure to group words into semantic categories based on their shared meanings. A detailed description of the third step can be found in section 3.4.

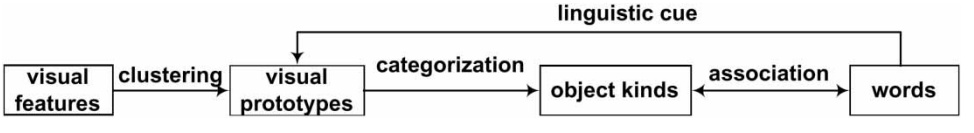

Figure 2. The overview of our approach.

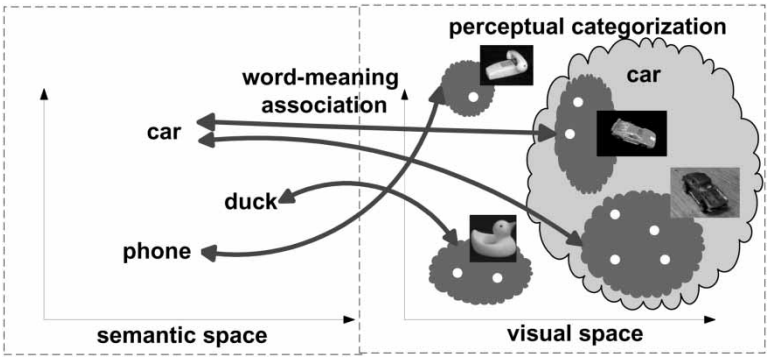

Figure 3. Object categorization and word–meaning association bootstrap each other to improve the overall performance. For example, the instances of the object 'red car' and the object 'green car' are initially clustered into two prototypes based on their visual similarities. Since both prototypes are likely to associate with the same linguistic label—the word 'car'—they will be grouped into one category that corresponds to the object 'car'.

Figure 3 illustrates a concrete example. Assume that there are several types (prototypes) of the object kind 'car' varying in colour and shape. For each specific type, there are several visual entities of that type obtained from different viewpoints, and maybe under different illuminations. At the first step, visual features are extracted from those entities to form feature vector representations. Then those feature vectors are clustered into several prototypes based on their similarities, each of which presumably corresponds to one type of the object kind 'car'. Note that this clustering step is relatively easy compared with an attempt to categorize all instances of different types of car into one category. Next, to the extent that we obtain visual prototypes and use them as visually grounded meanings of words, association probabilities between words and grounded meanings can be estimated. At the final step, those visual prototypes are iteratively merged into a set of categories, each of which corresponds to an object kind. The merging process is based not only on feature similarities in the feature space, but also on potential teaching signals provided by language. As shown in figure 3, the 'red car' prototype and the 'blue car' prototype might not be very similar in the feature space, but they will be merged into the same category because both of them are likely to be associated with the same linguistic label—the word 'car'.

3.2 *Prelinguistic object categorization*

Visual objects are represented by a set of colour, shape and texture features (Yu and Ballard 2004). Based on the work of Mel (1997), we constructed the visual features of objects that are large in number, invariant to different viewpoints and driven by multiple visual cues. Specifically, 64-dimensional colour features were extracted by a colour indexing method (Swain and Ballard 1991), and 48-dimensional shape features were represented by calculating histograms of local shape properties (Schiele and Crowley 2000). Gabor filters with three scales and five orientations were applied to the image. It was assumed that the local texture regions were spatially homogeneous, and the mean and the standard deviation of the magnitude of the transform coefficients were used to represent an object in a 48-dimensional texture feature vector. Thus, feature representations consisting of a total of 160 dimensions were formed by combining colour, shape and texture features, which provided fundamental advantages for fast, inexpensive recognition. Most classification algorithms, however, do not work efficiently in high-dimensional spaces because of the inherent sparsity of the data. This problem has traditionally been referred to as the curse of dimensionality. In our system, the 160-dimensional

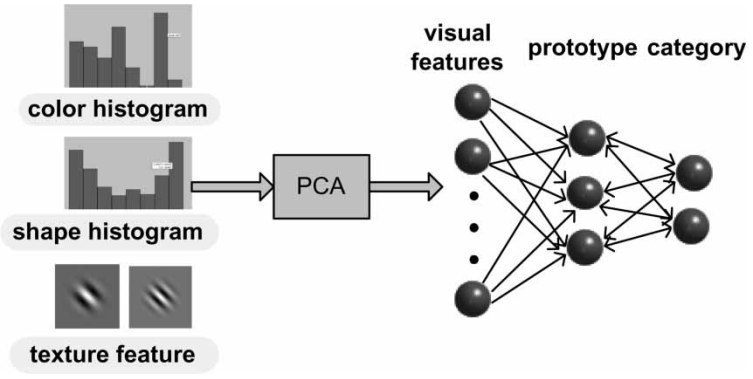

Figure 4. Visual features are extracted from the visual appearances of objects. After PCA, those features are used to cluster data into visual prototypes. The grouping of those prototypes forms object categories.

feature vectors were reduced into 15 dimensions by using principal component analysis (PCA) (Aggarwal and Yu 2000), which represents the data in a lower dimensional subspace by pruning away those dimensions that result in the least loss of information.

Figure 4 shows the basic architecture of object categorization. Each input component encodes a single feature dimension, with the activation of the component indicating the value of the stimulus on that dimension. We apply Gaussian mixture models to cluster perceptual features. Hence, each prototype component corresponds to a Gaussian in a multidimensional feature space. For a given feature vector $\mathbf{v}$, each prototype component is activated according to both the prior probability of this Gaussian prototype $p(\alpha)$ and the perceptual similarity of the feature vector to the αth Gaussian prototype $p(\mathbf{v}|\alpha)$, which can be specified by a probability distribution:

$$p(\mathbf{v}|\alpha) = \frac{1}{(2\pi)^{d/2}\det(\Sigma_\alpha)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{v}-m_\alpha)\Sigma_\alpha^{-1}(\mathbf{v}-m_\alpha)^{\mathrm{T}}\right),$$

where m_α and Σ_α are parameters of the αth Gaussian. We also assume the independence of the features and then enforce a block diagonal structure for the covariance matrix to capture the most important dependencies. In this way, the activation of a prototype component can be denoted as $p(\alpha)p(\mathbf{v}|\alpha)$, which is the probability of classifying the given feature $\mathbf{v}$ into the hidden prototype α; and the overall activation generated by a stimulus $\mathbf{v}$ can be expressed as $p(\mathbf{v}) = \sum_{\alpha=1}^{K} p(\alpha)p(\mathbf{v}|\alpha)$, where K is the number of prototypes. The standard expectation-maximization (EM) algorithm for mixture Gaussians is used to estimate the parameters of Gaussians as well as prior probabilities $p(\alpha)$.

Each prototype component is connected to an output component that correspond to a category. The probabilistic connection from a prototype component to a category component is determined by both perceptual similarity and co-occurring linguistic labels. The learning algorithm of estimating probabilistic links between prototype and category components is presented in section 3.4.

3.3 *Early word acquisition as statistical associative learning*

Next, each visual prototype is initially assigned to a distinct category (as shown in figure 5(a)). Then we can estimate the initial association probabilities between word and category components. Now we have a set of words $W = \{w_1, w_2, \ldots, w_N\}$ and a set of object categories $O = \{o_1, o_2, \ldots, o_M\}$, where N and M are the number of words and the number of object categories, respectively. These components are bi-directionally connected with each other,

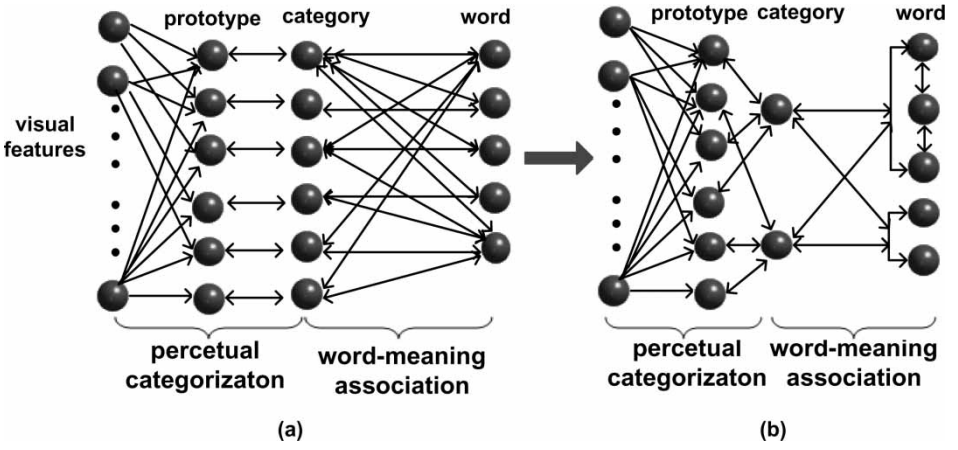

Figure 5. (a) Initially, we treat visual prototypes as object categories, and the probabilistic links between word components and category components are estimated. (b) The training algorithm estimates the probabilistic links between category and prototype components based on both associated linguistic labels and perceptual similarities. New categories lead to re-estimating the probabilistic links between word components and category components. In this way, the model iteratively estimates the probabilistic links between words and categories as well as the probabilistic links between categories and visual prototypes.

indicating that words and object categories are associated. To build lexicons, we want to estimate the association probabilities of word–object links.

The training data for word acquisition can be represented by a set $\chi = \{(S_w^l, S_o^l), 1 \leq l \leq L\}$, consisting of L learning situations. The lth situation includes two data streams (S_w^l, S_o^l). Each word stream S_w^l consists of n_l words $w_{u(1)}, w_{u(2)}, \ldots, w_{u(n_l)}$, and $u(i)$ can be selected from 1 to N, which corresponds to different words. Similarly, the extralinguistic object stream S_o^l includes m_l possible meanings $o_{v(1)}, o_{v(2)}, \ldots, o_{v(m_l)}$ and the value of $v(j)$ is from 1 to M, which corresponds to different objects. We can express the likelihood of observing the data set in terms of association probabilities between words and objects:

$$P(S_o^1, S_o^2, \ldots, S_o^L \mid S_w^1, S_w^2, \ldots, S_w^L) = \prod_{l=1}^{L} \sum_{a} p(S_o^l, a | S_w^l)$$
$$= \prod_{l=1}^{L} \frac{\epsilon}{(n_l+1)^{m_l}} \prod_{j=1}^{m_1} \sum_{i=0}^{n_1} p(o_{v(j)} | w_{u(i)}),$$

where the alignment a indicates which word is aligned with which object. $p(o_{v(j)}|w_{u(i)})$ is the association probability of a word–meaning pair and ϵ is a small constant.

The objective now is to choose those association probabilities (hidden variables) so that the likelihood function can be maximized, which can be accomplished by applying an EM algorithm. Informally, the EM algorithm starts with randomly assigned values of association probabilities, and then iteratively alternates two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, it computes the expected likelihood of generating the observation data given the current estimates of association probabilities. More specifically, the expected number of times that a particular word w_n in the word stream S_w^l generates a specific meaning o_m in the co-occurring object stream S_o^l is given by:

$$c(o_m | w_n, S_o^l, S_w^l) = \frac{p(o_m|w_n)}{p(o_m|w_{u(1)}) + \cdots + p(o_m|w_{u(n_1)})} \times \sum_{j=1}^{m_1} \delta(o_m, v(j)) \sum_{i=1}^{n_1} \delta(w_n, u(i)),$$

where δ is equal to one when both of its arguments are the same and equal to zero otherwise. In the M-step, the algorithm re-estimates those probabilities by maximizing the likelihood function:

$$p(o_m|w_n) = \frac{\sum_{l=1}^{L} c(o_m|w_n, S_o^l, S_w^l)}{\sum_{m=1}^{M}\sum_{l=1}^{L} c(o_m|w_n, S_o^l, S_w^l)}.$$

Once we have a new set of association probabilities, we can repeat the E-step and the M-step. This process continues until the likelihood function converges. The technical details of our method can be found in Yu and Ballard (2004).

3.4 *The emergence of object categorization and word–object association*

Now that we have not only words and visual prototypes (obtained from section 3.2) but also probabilistic links between them (obtained from section 3.3), we use these links to guide the categorization of visual prototypes, as shown in figure 5. In addition to this, the model is also able to group words into semantic categories based on their perceptually grounded meanings.

For a given set of co-occurrence data consisting of words and visual features of objects

$$\begin{aligned} S &= \{(S_w^1, S_v^1), (S_w^2, S_v^2), \ldots, (S_w^L, S_v^L)\} \\ &= \{(w_1^1, w_2^1, \ldots, w_{n_1}^1, v_1^1, v_2^1, \ldots, v_{m_1}^1), (w_1^2, w_2^2, \ldots, w_{n_2}^2, v_1^2, v_2^2, \ldots, v_{m_2}^2), \ldots, \\ &\quad (w_1^L, w_2^L, \ldots, w_{n_L}^L, v_1^L, v_2^L, \ldots, v_{m_L}^L)\} \end{aligned}$$

we seek to find a way to categorize visual prototypes into object kinds. Based on the categorization, we also want to calculate association probabilities between words and those object categories. The objective function to maximize the joint likelihood of the co-occurrences of words and visual features is as follows:

$$\arg\max_{O} \prod_{l=1}^{L} p(S_v^l, S_w^l|O),$$

where S_v^l is a set of visual features and S_w^l is a set of co-occurring words in the lth learning situation. O is a categorization of visual prototypes. We can rewrite the above as two parts:

$$p(S_v^l, S_w^l|O) = p(S_w^l|S_v^l, O)p(S_v^l|O) = p(S_w^l|O)\prod_{i=1}^{m_l}\sum_{\alpha=1}^{K} p(v_i^l|\alpha)p(\alpha|o_i^l),$$

where the first part is simplified as $p(S_w^l|O)$ because we assume that words S_w^l and visual features S_v^l are independent given the categorization of those visual representations. Moreover, $p(S_w^l|O)$ can be represented in terms of association probabilities of word–meaning pairs, which is described in section 3.3. The second part $p(S_v^l|O)$ corresponds to the categorization module where $p(v_i^l|\alpha)$ represents the probabilistic links between K prototypes and visual features, the values of which have already been obtained in previous steps (see section 3.2). $p(\alpha|o_i^l)$ represents the probabilistic links between category components and prototype components, which need to be estimated in this integrative step.

We initially assign each visual prototype α_i as a separate category o_i. Thus, the centroid of a category o_i is the same with a prototype α_i. The link between the jth object component and the ith prototype component can then be represented as $p(\alpha_i|o_j) = p(\alpha_i|o_j)/[\sum_k p(\alpha_i|o_k)]$. A greedy algorithm is then applied to merge incrementally category components in the set O to build a new set of categories O'. The merging decision is based on both perceptual similarities

and potential linguistic labels provided through word components. More specifically, the net change of merging two prototypes $d(O', O)$ includes two parts: (1) the change in visual space $d_{\text{within}}(O', O)$, and (2) the influence on the word-to-world mapping $d_{\text{between}}(O', O)$. In practice, we add a weight factor to balance the effects from these two sources.

Assume that we want to merge two category components o_p and o_q. Thus, the only difference between O' and O is that o_p and o_q are assigned to a new category o_m in O'. The perceptual differences between these two object categories are measured by:

$$d_{\text{within}}(O', O) = \sum_{l=1}^{L} \sum_{i=1}^{m_l} \log \frac{\sum_{\alpha=1}^{K} p(v_i^l|\alpha) p(\alpha|o_p)}{\sum_{\alpha=1}^{K} p(v_i^l|\alpha) p(\alpha|o_q)}.$$

Meanwhile, the influence of merging two object categories on word–meaning association can be estimated from:

$$d_{\text{between}}(O', O) = \sum_{l=1}^{L} \log \frac{p(S_w^l|O')}{p(S_w^l|O)}.$$

To estimate $p(S_w|O')$, we would in principle need to know new association probabilities between words and new object kinds. To reduce the computational load, we assume that the merging of categories o_p and o_q will not affect association probabilities of categories other than o_p and o_q. Thus, $p(w_i|o_j)$ will remain unchanged for $j \neq p$ and $j \neq q$. For the new object category o_m in O', its association probability to a specific word w_i can be approximated by:

$$p(w_i|o_m) \approx p(w_i|o_p)p(o_p|o_m) + p(w_i|o_q)p(o_q|o_m) = \frac{p(w_i|o_p)p(o_p) + p(w_i|o_q)p(o_q)}{p(o_p) + p(o_q)}.$$

The net change $d(O', O)$ is then used to determine whether two categories o_p and o_q will be merged. The merging decision will lead to updating not only the probabilistic labels between prototype and category components, but also the probabilistic links between words and object categories. As shown in algorithm 1, our method runs multiple times and allows for merging visual prototypes progressively.

Algorithm 1. *Categorization and Association.*
Cluster visual features using Gaussian mixture models, which form the initial set of category components O.
Compute the association probabilities $p(w_i|o_j)$ between category components and word components.
Repeat the following steps until no more merging operations needed:
repeat
(1) Find n closest pairs of object categories from O as candidate pairs based on the similarities of visual features, and sort them in a list based on similarity scores.
(2) Remove from the sorted list all the candidate pair (o_p, o_q) where either o_p or o_q is part of another pair higher in the list.
(3) Compute $d(O', O)$ of all candidate pairs.
(4) Go through every pair in the list and merge object categories (o_p, o_q) into one object category o_m if the net change of merging is less than a pre-defined value.
(5) Replace O with O' and compute new association probabilities $p(w|o)$.
until converge.

We also apply the same mechanism to group words into several categories (bags of words) that share the same object category; and we run the categorization of visual objects and the grouping

of words in parallel on alternate iterations. Compared with other algorithms, this approach estimates association probabilities of word–meaning pairs as well as categorizes visual objects in an interactive manner. As a result, the cross-modal information effects object categorization by guiding the grouping of visual prototypes. Consequently, a new set of categories also changes the association probabilities between word and category components.

4. Experiments

Subjects wore a head-mounted CCD camera to capture a first-person point of view. Visual data were collected at a resolution of 320 columns by 240 rows of pixels. Acoustic signals were recorded using a headset microphone at a rate of 22 kHz with 16-bit resolution. Six subjects, all native speakers of English, participated in the experiment. They were asked to narrate the picture book 'I went walking' in English. The book is for 1–3-year-old children and the story is about a young child who goes for a walk and encounters several familiar friendly animals. For each page of the book, there are several animals, as shown in figure 6. Subjects were instructed to pretend that they were telling this story to a child so that they should keep verbal descriptions of pictures as simple and clear as possible. We collected multisensory data when they performed the task, which were used as the training data for our computational model.

The regions of visual objects were segmented from the background based on the method described in Yu and Ballard (2004). For each region, we applied the visual feature extraction method described in section 3.2. For acoustic signals, we implemented an endpoint detection algorithm to segment the speech stream into spoken utterances. Each spoken utterance contained one or more spoken words. Then the 'Dragon Naturally Speaking' software was employed to convert spoken utterances into text. Given a spoken utterance consisting of several spoken words, the speech recognizer converted the continuous wave pattern into a series of recognized words by considering phonetic likelihoods and grammars. In practice, the recognition rate was above 90% in our experiments. Next, we paired visual features of objects with the text of individual spoken utterances to form co-occurring word–context parallel sequences.

By applying the learning mechanism described in section 3, the present model obtained grounded lexical items, including categories of visual objects paired with their corresponding linguistic labels. There were 12 animals and some other objects (e.g. tree, basket) shown in the picture book. Compared with animals, those objects rarely occur so are barely mentioned in speech. Consequently, we did not treat them as lexical items that the model was expected to acquire. To evaluate the experimental results, we defined the following three measures: (1) *Word–meaning grounding accuracy* measures the percentage of the perceptual

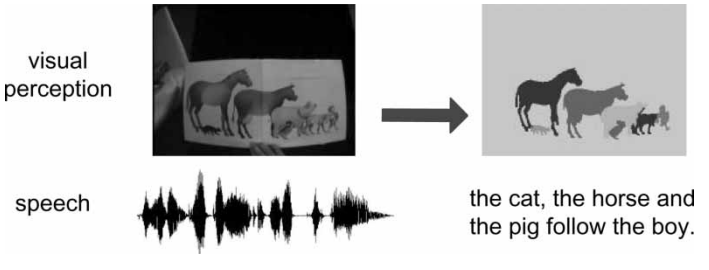

Figure 6. Left: spoken descriptions and image sequences are collected when subjects narrate the picture book. Right: visual objects are segmented from the background scene and then multiple features are extracted to form perceptual representations. The acoustic signals are converted into text. Those visual features and words are used as the training data.

representations of objects that are correctly associated with linguistic labels. Given a visual feature, the model is expected to find related linguistic labels. Thus, the corresponding word components will be activated through their connections with category components. Likewise, given a word, the model is able to activate the corresponding components of visual prototypes that represent perceptually grounded meanings. (2) *Word spotting accuracy* measures the percentage of word–object pairs that are spotted by the computational model. This measure provides a quantitative indication about the percentage of grounded lexical items that can be successfully found. (3) *Object categorization accuracy* measures the percentage of visual features that are grouped into correct object categories.

To demonstrate more directly the role of links between perceptual categorization and linguistic development, we included two other learning conditions in this experiment. In a no-word condition (similar to the one in Waxman and Hall (1993)), we applied Gaussian Mixture Models to cluster visual features. The clustering results were used as distinct object categories and directly associated with words. Thus, there was no interaction between association and categorization in this method, which clearly reduced the amount of computation needed. Note that the above two learning conditions are intended to simulate closely human development in a natural environment. Therefore, there are no explicit linguistic labels provided. In the third condition—the labelling condition (similar to the linguistic-cued condition in Yoshida and Smith (2005))—we assume that there is a linguistic label paired with every instance of an object. Thus, there is no ambiguity in the word-to-world mapping and every word is perfectly mapped to its meaning. This is, of course, an ideal learning condition that does not exist in a natural learning environment. The purpose here is to provide a complete evaluation of the influence of language in object categorization. The results in this condition indicate to what degree linguistic labels can enhance object categorization. In practice, we applied support vector machines (SVM) on visual data that are paired linguistic labels. With explicit teaching signals and the start-of-art classification method, this approach provides a baseline and illustrates the difficulty of categorizing visual data in the experiment. Table 1 shows the comparative results of these three learning conditions.

The significant difference between the first two conditions lies in the fact that there exists a multitude of co-occurring word-to-object pairs in natural environments that infants are situated in, and modelling the links between language and category learning captures cross-modal and spatio-temporal constraints in co-occurring multimodal data. This fact shows that language can strengthen learning about object categories. It is not surprising that the labelling condition performs best with additional teaching labels of objects provided during the training. The categorization accuracy obtained from our model, however, is closer to that of a supervised labelling condition. Both learning conditions use linguistic labels to help object categorization while correlated linguistic cues are provided in the labelling condition but not in the link condition. Nonetheless, the model in the link condition simulates the possible correlations between object name learning and object categorization. By doing so, it is able to bootstrap these two advances and form a developmental feedback loop, showing that these two difficult learning problems can be fundamentally simplified by considering them in a general system.

Table 1. Results of object categorization and word learning.

	Word–meaning association (%)	Word spotting (%)	Object categorization (%)
Link condition	76.6	73.2	82.9
No-word condition	51.9	52.5	62.3
Labelling condition	82.1	77.6	85.3

Most importantly, this observation is quite in line with the results obtained from human subjects (Waxman 2004, Jones and Smith 2004, Yoshida and Smith 2005), suggesting not only that our model is cognitively plausible, but also that the emergence of links between perceptual and linguistic development can be appreciated by both human learners and the computational model.

5. Discussion and conclusion

Computational models of development and cognition have changed radically in recent years. Many cognitive scientists have recognized that models that incorporate constraints from embodiment—that is, how mental and behavioural development depends on complex interactions among brain, body and environment (Clark 1997)—are more successful than models that ignore these factors. Language represents perhaps the most sophisticated cognitive system acquired by human learners, and it clearly involves complex interactions between a child's innate capacities and the social, cognitive and linguistic information provided by the environment (Gleitman and Newport 1995). The model outlined in this study focuses on the initial stages of word learning – How are words attached to their perceptually grounded meanings? Many existing models of language acquisition have been evaluated by artificially derived data of speech and semantics (Brent and Cartwright 1996, Siskind 1996, Bailey *et al.* 1998, Cohen *et al.* 2001). In contrast, multisensory data (materials used by the model) are real and natural. Our model proved successful by taking advantage of recent advances in cognitive modelling, machine learning and computer vision.

There is evidence that, from an early age, infants are able to form perceptually-based category representations (Quinn *et al.* 1993). Those categories are highlighted by the use of common words to refer to them. Thus, the meaning of the word 'dog' corresponds to the category of dogs, which is initially a non-linguistic mental representation in the brain. Furthermore, Schyns and Rodet (1997) argued that the representations of object categories emerge from the features that are perceptually learned from visual input during the developmental course of object recognition and categorization. In this way, object naming by young children is essentially about mapping words to selected perceptual properties. Our model simulates the underlying mechanism of this mapping process. It represents the meanings of object names as visual features consisting of shape, colour and texture extracted from the visual appearances of objects. Then the model learns to associate those visually grounded meanings with linguistic labels.

Moreover, our model attempts to answer a more intriguing question in object naming – How does word learning during infancy influence the formation of the categories to which those words refer? Waxman and Hall (1993) showed that linguistic labels may facilitate the formation of a category by infants at 16–21 months of age. Hence, the formation of categories in the presence of auditory input is possibly based on both the similarities of perceptual features and on shared linguistic labels. Our current model offers a formal account of how linguistic, perceptual and conceptual advances are linked. The model demonstrates the bi-directional relationship between lexical acquisition and object category learning. Thus, object categorization provides mental representations of meanings that are mapped to words. Meanwhile, linguistic labels enhance object categorization by providing additional teaching signals. Our simulation results provide a dynamic view of the essential interactions between word learning and object categorization that is quite in line with evidence obtained from empirical studies (Roberts and Jacob 1991, Waxman 2004, Landau 2004, Yoshida and Smith 2005), suggesting the cognitive plausibility of our model.

Several aspects in word acquisition are simplified in the present model. First, the model uses transcriptions of speech as linguistic input but not raw speech signals. Segmenting continuous speech into isolated words is the first task in language acquisition. Indeed, English-learning infants do display some ability to segment words as early as about 7.5 months (Jusczyk and Aslin 1995). This model focuses on how they acquire the meanings of these segmented words. Second, the model is a passive learner that perceives multimodal data without active exploration and manipulation of the environment, while a true embodied agent has the opportunity to explore and acquire information more effectively than the passive one here (Clark 1997, Rizzolatti and Arbib 1998). Third, Goldstone *et al.* (2001) showed that judgements of the similarity of two members of the same category agree more to a third non-categorized object as a result of category learning. This suggests that grouping two objects together changes their internal descriptions because features are created to subserve the representations and categorizations of objects (Schyns *et al.* 1998). More specifically, the elements that two objects share become more important parts of the objects' descriptions. In the context of our model, this suggests that language could potentially influence not only the computation in the categorization module, but also that in the perceptual module (see figure 5). Further studies are needed to explore whether language shapes both perceptual learning and conceptual learning simultaneously and, if so, how these two learning processes interact and co-ordinate over the course of acquisition.

The principal technical contribution of this paper is to use multimodal data for object categorization. Clustering visual data is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes that can be high-dimensional. The curse of dimensionality in high-dimensional clustering is well known, and researchers in statistics, pattern recognition and machine learning have developed various algorithms to tackle it. The typical technique is to transform the original feature space into a lower dimensional space. Considering the technical difficulty of the problem, it is surprising that even young children can easily recognize various objects. How do they accomplish this? We attempt to simulate the underlying mechanism applied by human learners. Our model explores the possibility of utilizing the spatio-temporal and cross-modal constraints in multimodal data for perceptual categorization. The results indicate that the unimodal categorization problem can be fundamentally simplified by incorporating linguistic cues co-occurring in a natural environment. Furthermore, the model considers perceptual categorization and word–meaning mapping in an integrative framework, suggesting a unified view of different aspects of human development and learning. Hence, this work potentially provides some useful insights for further experimental studies about how the brain co-ordinates linguistic and perceptual learning, and combines information from different sensory modalities.

Acknowledgements

I wish to express my thanks to Dana H. Ballard, Robert A. Jacobs and Rob Goldstone for invaluable discussions of ideas contained in the paper. I am grateful to Susan Jones and anonymous reviewers for thoughtful comments on an earlier draft of this paper.

References

C.C. Aggarwal and P.S. Yu, "Finding generalized projected clusters in high dimensional spaces", in *Sigmod*, Dallas, Texas, USA: ACM, 2000.

D. Bailey, N. Chang, J. Feldman and S. Narayanan, "Extending embodied lexical development", in *Proceedings of the 20th Annual Meeting of the Cognitive Science Society Cogsci-98*, 1998.

S. Becker, "Mutual information maximization: models of cortical self-organization", *Network: Comput. Neural Syst.*, 7, pp. 7–31, 1996.

G. Behl-Chadha, "Basic-level and superordinate-like categorical representations in early infancy", *Cognition*, 60, pp. 105–141, 1996.
M.R. Brent, "Speech segmentation and word discovery: a computational perspective", *Trends Cognitive Sci.*, 3, pp. 294–301, 1999.
M.R. Brent and T.A. Cartwright, "Distributional regularity and phonotactic constraints are useful for segmentation", *Cognition*, 61, pp. 93–125, 1996.
A. Cangelosi, A. Greco and S. Harnad, "From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories", *Connection Sci.*, 12, pp. 143–162, 2000.
A. Cangelosi and D. Parisi, "The emergence of a 'language' in an evolving population of neural networks", *Connection Sci.*, 10, pp. 83–97, 1998.
A. Clark, *Being There: Putting Brain, Body and World Together Again*, Cambridge, MA: MIT Press, 1997.
E. Clark, "What's in a word? On the child's acquisition of semantics in his first language", in *Cognitive Development and the Acquisition of Language*, T.E. Moore, Ed., New York: Academic Press, 1973.
P.R. Cohen, T. Oates, N. Adams and C.R. Beal, "Robot baby 2001", in *Lecture Notes in Artificial Intelligence*, Vol 2225, Washington, DC, Springer-Verlag: London, 2001, pp. 32–56.
V.R. de Sa and D. Ballard, "Category learning through multimodality sensing", *Neural Comput.*, 10, pp. 1097–1117, 1998.
J. Elman, E. Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi and K. Plunkett, *Rethinking Innateness: A Connectionist Perspective on Development*, Cambridge, MA: MIT Press, 1996.
D. Gentner, "A study of early word meaning using artificial objects: What looks like a jiggy but acts like a zimbo?", in *Readings in Developmental Psychology*, 2nd edn, J. Gardner, Ed., Boston: Little Brown, 1978, pp. 137–142.
J. Gillette, H. Gleitman, L. Gleitman and A. Lederer, "Human simulations of vocabulary learning", *Cognition*, 73, pp. 135–176, 1999.
L. Gleitman, "The structural sources of verb meanings", *Language Acquisition*, 1, pp. 1–55, 1990.
L. Gleitman and E. Newport, "The invention of language by children: environmental and biological influences on the acquisition of language", in *An Invitation to Cognitive Science: Language*, VOL 1, L. Gleitman and Liberman, Eds, Cambridge, MA: MIT Press, 1995.
R. Goldstone, Y. Lippa and R. Shiffrin, "Altering object representations through category learning", *Cognition*, 78, pp. 27–43. 2001.
S. Harnad, "The symbol grounding problem", *Physica D*, 42, pp. 335–346, 1990.
S.S. Jones and L.B. Smith, "The effects of object name learning on object perception: a deficit in late talkers", *J. Child Language*, 31, pp. 1–18, 2004.
P.W. Jusczyk and R.N. Aslin, "Infants detection of the sound patterns of words in fluent speech", *Cognitive Psychol.*, 29, pp. 1–23, 1995.
B. Landau, "Perceptual units and their mapping with language: how children can (or can't?) use perception to learn words", in *Weaving a Lexcon*, G. Hall and S.R. Waxman, Eds, Cambridge, MA: MIT Press, 2004.
B. Landau, L. Smith and S. Jones, "Object shape, object function, and object name", *J. Mem. Language*, 36, pp. 1–27, 1997.
B. Landau, L. Smith and S. Jones, "Object perception and object naming in early development", *Trends Cognitive Sci.*, 2, pp. 19–24, 1998.
M. Lungarella, G. Metta, R. Pfeifer and G. Sandini, "Developmental robotics: a survey", *Connection Sci.*, 2004.
B. MacWhinney, *Linguistic Categorization*, Benjamins: New York, 1989, pp. 195–242.
D. Mareschal and P. Quinn, "Categorization in infancy", *Trends Cognitive Sci.*, 5, pp. 443–450, 2001.
H. McGurk and J. MacDonald, "Hearing lips and seeing voices", *Nature*, 264, pp. 746–748, 1976.
B.W. Mel, "Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition", *Neural Comput.*, 9, pp. 777–804, 1997.
K. Plunkett, C. Sinha, M.F. Miller and O. Strandsby, "Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net", *Connection Sci.*, 4, pp. 293–312, 1997.
P. Quinn, P. Eimas and S. Rosenkrantz, "Evidence for representations of perceptually similar natural categories by 3-month old and 4-month old infants", *Perception*, 22, pp. 463–375, 1993.
P. Quinn, P. Eimas and M. Tarr, "Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants", *J. Exp. Child Psychol.*, 79, pp. 78–94, 2001.
T. Regier, *The Human Semantic Potential: Spatial Language and Constrained Connectionism*, Cambridge, MA: MIT Press, 1996.
T. Regier, "Emergent constraints on word-learning: a computational review", *Trends Cognitive Sci.*, 7, pp. 263–268, 2003.
G. Rizzolatti and M.A. Arbib, "Language within our grasp", *Trends Neurosci.*, 21, pp. 188–194, 1998.
K. Roberts and M. Jacob, "Linguistic versus attentional influences on nonlinguistic categorization in 15-month-old infants", *Cognitive Dev.*, 6, pp. 355–375, 1991.
D. Roy and A. Pentland, "Learning words from sights and sounds: a computational model", *Cognitive Sci.*, 26, pp. 113–146, 2002.
J.R. Saffran, E.L. Newport and R.N. Aslin, "Word segmentation: the role of distributional cues", *J. Mem. Language*, 35, pp. 606–621, 1996.
B. Schiele and J.L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms", *Int. J. Comput. Vision*, 36, pp. 31–50, 2000.

P. Schyns and L. Rodet, "Categorization creates functional features", *J. Exp. Psychol.: Learn. Mem. Cognit.*, 23, pp. 681–696, 1997.
P.G. Schyns, R.L. Goldstone and J.-P. Thibaut, "The development of features in object concepts", *Behav. Brain Sci.*, 21, pp. 1–54, 1998.
J.M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings", *Cognition*, 61, pp. 39–61, 1996.
V.M. Sloutsky and Y.-F. Lo, "How much does a shared name make things similar? Part: linguistic labels and the development of similarity judgment", *Dev. Psychol.*, 35, 1999.
L. Smith, "How to learn words: an associative crane", in *Breaking the Word Learning Barrier*, R. Golinkoff and K. Hirsh-Pasek, Eds, Oxford: Oxford University Press, 2000, pp. 51–80.
L. Smith and D. Heise, "Infants' use of textural information to discriminate between artifacts and natural kinds". PhD thesis, Indiana University (2000).
L. Smith, S. Jones and B. Landau, "Naming in young children: a dumb attentional mechanism?", *Cognition*, 60 (2), pp. 143–171, 1996.
L. Smith, S. Jones, B. Landau, L. Gershkoff-Stowe and L. Samuelson, "Object name learning provides on-the-job training for attention", *Psychol. Sci.*, 13, pp. 13–19, 2002.
L.B. Smith, "Learning to recognize objects", *Psychol. Sci.*, 14 (3), pp. 244–250, 2003.
E. Spelke, G. Gutheil and G.V.D. Walle, "The development of object perception", in *An Invitation to Cognitive Science; Vol. 2 Visual Cognition*, D.N. Osherson and S.M. Kosslyn, Eds, Cambridge, MA: MIT Press, 1995.
L. Steels, "Synthesizing the origins of language and meanings from co-evolution", in *Evolution of Human Language*, J. Hurford, Ed., Edinburgh: Edinburgh University Press, 1997.
L. Steels and P. Vogt, "Grounding adaptive language game in robotic agents", in *Proceedings of the 4th European Conference on Artificial Life*, C. Husbands and I. Harvey, Eds, London: MIT Press, 1997.
M.J. Swain and D. Ballard, "Color indexing", *Int. J. Comput. Vision*, 7, pp. 11–32, 1991.
J. Tenenbaum and F. Xu, "Word learning as Bayesian inference", in *Proceedings of the 22nd Annual Conference of Cognitive Science Society*, L. Gleitman and A. Joshi, Eds, Mahwah, NJ: ErlBaum, 2000, pp. 517–522.
S. Waxman and D. Hall, "The development of a linkage between count nouns and object categories: evidence from fifteen- to twenty-one-month-old-infants", *Child Dev.*, 29, pp. 257–302, 1993.
S.R. Waxman, "Everything had a name, and each name gave birth to a new thought: links between early word learning and conceptual organization", in *Weaving a Lexcon*, G. Hall and S.R. Waxman, Eds, Cambridge, MA: MIT Press, 2004.
S.R. Waxman and D.B. Markow, "Words as invitations to form categories: evidence from 12- to 13-month-old infants", *Cognitive Psychol.*, 29, pp. 257–302, 1995.
J. Weng, Y. Zhang and Y. Chen, "Developing early senses about the world: 'object permanence' and visuoauditory real-time learning", in *Proceedings of the International Joint Conference on Neural Networks*, 2003, pp. 2710–2715.
F. Xu, "The role of language in acquiring object kind concepts in infancy", *Cognition*, 85, pp. 223–250, 2002.
H. Yoshida and L.B. Smith, "Linguistic cues enhance the learning of perceptual cues", *Psychol. Sci.*, 16 (2), 2005.
C. Yu and D.H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions", *ACM Trans. Appl. Percept.*, 1, pp. 57–80, 2004.
C. Yu, D.H. Ballard and R.N. Aslin, "The role of embodied intention in early lexical acquisition", in *Proceedings of the 25th Cognitive Science Society Annual Meetings*, 2003, pp. 1293–1298.