Modeling Cross-Modal Interactions in Early Word Learning

ARTICLE *in* IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT · DECEMBER 2013

Impact Factor: 1.35 · DOI: 10.1109/TAMD.2013.2264858

2 AUTHORS:

Nadja Althaus
University of Oxford
7 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Denis Mareschal
Birkbeck, University of London
148 PUBLICATIONS **1,694** CITATIONS
SEE PROFILE

Modeling Cross-Modal Interactions in Early Word Learning

Nadja Althaus and Denis Mareschal

***Abstract*—Infancy research demonstrating a facilitation of visual category formation in the presence of verbal labels suggests that infants' object categories and words develop interactively. This contrasts with the notion that words are simply mapped "onto" previously existing categories. To investigate the computational foundations of a system in which word and object categories develop simultaneously and in an interactive fashion, we present a model of word learning based on interacting self-organizing maps that represent the auditory and visual modalities, respectively. While other models of lexical development have employed similar dual-map architectures, our model uses active Hebbian connections to propagate activation between the visual and auditory maps during learning. Our results show that categorical perception emerges from these early audio–visual interactions in both domains. We argue that the learning mechanism introduced in our model could play a role in the facilitation of infants' categorization through verbal labeling.**

***Index Terms*—Categorization, computational modeling, cross-modal interactions, self-organizing maps, word learning.**

I. Introduction

At the start of lexical development, infants face the task of making links between words and categories of objects. This is not a simple task, considering that both their language and category systems are still immature by the time the first words are produced (around 12 months of age). While it seems intuitive that infants only engage in learning words they can readily map onto preexisting (i.e., already formed) concepts, evidence from studies investigating the impact of labeling on category formation suggests otherwise: it appears that, at least to some extent, hearing similar labels for objects may alter the way categories are learned. For this reason we propose a model of word learning that allows interactions between the visual and auditory domains from the start, with a growing impact of representation in each modality on the other.

Manuscript received August 07, 2012; revised November 19, 2012; accepted November 23, 2012. Date of publication June 28, 2013; date of current version December 09, 2013. This work was supported by the European Commission Grant MESTCT-2005-020725, the Economic and Social Research Council U.K. Grant RES-062-23-0819, and the Wellcome Trust Grant 084386/Z/07/Z . The work of N. Althaus was supported by the Winkler Career Development Fellowship in Experimental Psychology at St Hugh's College, University of Oxford. The work of D. Mareschal was supported by a Royal Society-Wolfson Research Merit Award.

N. Althaus is with the Department of Experimental Psychology, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: nadja.althaus@psy.ox.ac.uk).

D. Mareschal is with the Department of Psychological Science, Birkbeck, University of London, London WC1E 7HX, U.K. (e-mail: d.mareschal@bbk.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2013.2264858

A. Categorization and Labeling in Infancy

During their first year of life, infants develop a phonological system attuned to their native language [1]. Gradually they also develop segmentation skills in order to identify recurring units of speech (e.g., [2], [3]), and finally they must also be able to link those to real-world entities, whether this be objects they encounter (such as "ball." "teddy," "bottle") or more abstract concepts (e.g., "all gone"). Although infants begin recognizing words as early as 6 months after birth [4], referent selection is far from trivial: even after having identified that "ball" is a word, it must be determined whether it refers to the round, red object that daddy brought to play, to the round shape in general, the color red, to "bouncing" or perhaps even to the game "catch and throw" [5]. One of the questions therefore is how infants manage to learn words as referring to an object category; i.e., a set of, potentially similar, objects. It appears intuitive that infants might engage in categorization before the onset of word learning and, with increasing phonological and segmentation skills, map words onto preexisting nonverbal concepts. This is in line with literature reporting extensive categorization skills as early as 3 to 4 months of age (e.g.,[6]) which continue to develop throughout the first year of life [7]–[10].

However, this view is contradicted by a growing body of research that indicates that the presence of verbal labels can have an impact on category formation in experimental settings. For example, Waxman and Markow [11] familiarized 12- to 13-month-olds with a set of toy objects and tested categorization by subsequently presenting a novel within-category object together with an out-of-category object. While this test was always presented without a label, infants heard either labeling or nonlabeling phrases during familiarization. Those infants who had heard consistent, novel labels during familiarization showed a preference for the out-of-category object, suggesting successful category formation, whereas the infants hearing nonlabeling phrases did not exhibit a preference. The authors' interpretation was that labels had facilitated category formation, and suggested further that labels might "highlight commonalities" between exemplars. This type of facilitation effect has since been shown to be specific to consistent as opposed to variable labels [12], and to linguistic stimuli as opposed to, e.g., tone sequences [13]–[15]. Besides, this influence of words on categorization has been observed even for 3-month-olds, who are far from the age at which word learning – or at least fast mapping in laboratory settings – typically occurs [15]. While these studies focused on the facilitation of category formation in cases where infants were unsuccessful without labels, Plunkett *et al.* [16] have shown that labels may also modulate

category formation in the case where a visual stimulus set could be "parsed" into one large category or two subcategories depending on whether feature correlations were used as indicators for category membership. In this case, 10-month-olds formed two categories in silence or in the presence of two labels correlated with visual information, but formed a single category when all objects were presented with the same label. While this "merging" of categories leaves open whether labels did indeed play a constructive role (since the same result could have been achieved by the label "overshadowing" visual detail, cf. [17]), Althaus and Westermann [18] provided evidence that labels can also cause infants to split visual categories. In their study, infants perceived a visual morphing continuum as just one category in silence or when identical labels accompanied the pictures, but as soon as distinctive labels were provided that correlated with either end of the continuum infants divided the continuum into two separate categories. Althaus and Mareschal [19] further showed in an eye-tracking paradigm that labels modulate infants' eye movements during category learning, causing infants to pay more attention to low-variability object parts in the presence of labels—a finding consistent with Waxman and Markow's [11] hypothesis that labels highlight commonalities.

At the same time there is also some evidence for a disruptive impact of labels on category formation. Robinson and Sloutsky [17] conducted a study using very similar methods to those of [14] and [15]. However, they contrasted a label condition (familiarization with images of cats, accompanied by the label "cat") with an entirely silent condition and a synthesized-sound condition, rather than including a no-label condition (a condition including speech, but no novel labels, such as "Look at this one!"). Their results indicated that 12-month-olds in fact learned better in the absence of labels, while labels disrupted learning less than the synthesized sound. These results call into question whether labels in Waxman et al.'s studies were indeed facilitative since a silent control condition was not reported [16]. To further our understanding of the learning problem at the heart of this unresolved empirical phenomenon – do labels facilitate or hinder – we must first understand the computational challenges involved in interactions between word and category acquisition.

A model of word learning that represents not just the acquisition of an arbitrary mapping between symbols in two domains (words and objects) but also the relationship between representations in those domains (i.e., instances of object categories, as well as utterances of a word, for that matter) therefore needs to allow for such audiovisual interactions to take place. Of course, when exactly those interactions begin or whether they are indeed present from the onset of auditory and visual learning is somewhat of an open question, but the finding that there may even be differences in category formation with and without labels at 3 months [15] suggests that early interactions are a possibility.

What are the consequences of early interactions between visual and auditory domains? From an information-theoretic point of view, it is possible that this would lead to a bootstrapping effect, where the additional information present in one domain may aid learning in the other. However, is it beneficial for category and word representation to influence each other at a stage when, perhaps, both are immature? The main motivation for the interactive model described in this paper is to investigate this question.

B. Models of the Acquisition of Word-Object Mappings

A number of computational models have been proposed in the past to simulate learning at the interface between language and object processing. Several of these [20]–[22] have mainly provided evidence that the information available in the joint auditory and visual information streams infants typically perceive is sufficient for extracting word-object mappings, and that this multimodal information indeed allows better word segmentation or identification than the auditory stream alone. Yu [21] has further provided a computational account of the impact of labeling on categorization. In his model, labels help to bootstrap categories by providing a way of identifying (potentially dissimilar) visual exemplars through label-cooccurrence. However, all of these models employ computation-intensive statistical algorithms that need iterative access to the full data set, while at the same time a developmental trajectory of learning is not directly observed. As such, none of these approaches provide a plausible, mechanistic, developmental account of the interaction of labeling and categorization in infants.

Some connectionist approaches have gone beyond this by using learning mechanisms that correspond more closely to the sort of processing that may occur in a neural system. Schyns [23] modelled concept formation with a self-organizing map modified by the addition of a supervised naming mechanism. This way, the model integrated unsupervised self-organization processes with supervised learning, resulting in a prototype naming effect (the model produced the correct label faster for a category prototype than for exemplars distant from the category centroid). The model further demonstrated how hierarchical organization emerges in a top-down way (i.e., superordinate categories were learned first), and how differences in expertise in this context arose from different exposure. Plunkett *et al.* [24] approached the word learning problem with a connectionist simulation. Their model was an autoencoder receiving images and labels as input. Both types of data were initially processed in separate pathways, and projected onto separate sets of output units (for which the target was to reconstruct the input pattern). However, both pathways met in a layer of hidden units, which therefore served to encode the image/label association. The model reproduced several important aspects of language acquisition in infants. Learning proceeded in a nonlinear fashion, resembling the vocabulary "spurt" observed in infants. The model also exhibited comprehension-production asymmetries: producing the correct image-output upon presentation of a label appeared easier to learn for the model than producing the correct label upon presentation of an image pattern. Furthermore the authors found that this cross-modal network learned the image categories faster than a unimodal version which did not involve any labeling. Clearly, the model was able to exploit the additional information given by the second modality. A similar architecture was used by Schafer and Mareschal [25] to simulate word learning data from 8- and 14-month-olds. Stager

and Werker [26] interpreted infants' failure at 14 months to map a minimal pair of words ("dih", "bih") to distinct objects as a deficiency in phonological encoding due to the processing demands involved in word-object mapping. In a purely auditory discrimination task even 8-month-olds perform well on this contrast. Schafer and Mareschal employed connectionist modeling to demonstrate that the seemingly complex looking patterns might in fact be the outcome of an interaction of age (i.e., more experience, or longer training in the model) and inherent similarity structure in the stimuli. This was later confirmed experimentally, as 14-month-olds could be shown to learn the relevant word-object mappings using more sensitive methodology [27].

Li *et al.* (2004) [28] introduced DevLex, a model consisting of two interconnected growing self-organizing maps, one for lexical-semantic information, and one for phonological information. Hebbian connections were formed between the maps so as to connect units that were coactivated. Thus the network learned, over time, to associate word forms with their semantic representation. Learning in this model consisted of two different learning modes—Kohonen's [29] self-organizing map (SOM) algorithm, as well as a learning mode based on adaptive resonance theory (ART) [30]. While the SOM mode was responsible for map organization, the ART mode was used for recruiting new network units to simulate vocabulary growth. Essentially, the model started out by learning in SOM mode, optimising the topological organization of the existing units. This was followed by a gradual transition to ART mode, in which new units were recruited whenever the distance between input and map units failed to exceed a certain threshold. DevLex successfully simulated lexical confusion and age of acquisition effects (i.e., words learned early are processed more easily). Further, the model demonstrated that linguistic categories (nouns, verbs, adjectives and function words) can emerge in this form of learning, rather than having to be hard-wired into the system. An extension of this model was proposed by Li *et al.* (2007) [31]. Working with classic self-organizing maps instead of the growing maps, and including an additional "phonetic output map," this model was capable of simulating a vocabulary spurt as well as frequency effects. Mayor and Plunkett [32] introduced another dual-map model with Hebbian links. A visual map was trained with distortions of prototypical dot patterns. Simultaneously, an acoustic map was trained with speech samples. Model development included an early phase of synapto-genesis (i.e., increasing connectivity) and a later phase during which inactive Hebbian links were pruned away. While both maps essentially developed independently, "joint attentional events" trained the Hebbian links between them: here, an object pattern was presented simultaneously with a matching acoustic label. The model demonstrated successfully that a high amount of joint attentional activity is beneficial for vocabulary growth. The authors argued further that "taxonomic responding" (i.e., the mapping of a label to all exemplars of an object's category, rather than just to one exemplar) was an emergent property of the model. Taxonomic responding after just one object-label exposure was higher when the maps had developed independently for a longer period. This attributes a prominent role to the emergence of prelinguistic categorization without an early interaction with word learning.

A different approach, also using a self-organizing map, was taken up by Gliozzi *et al.* [33], who presented a model simulating the empirical results presented by Plunkett *et al.* [16], in which labels caused infants to merge separate categories if they were presented with a common label. In contrast to the multiple-map approaches discussed above, this model only uses a single map which receives input from the two sensory domains. This model was specifically aimed at simulating the influences of labeling on categorization, by processing the label as the equivalent of an additional feature. Categorization of the objects was measured by evaluating the distance between the test object and the best matching unit after presenting each training exemplar just once. The results mimicked the "merging" of visual categories that was found in the experiments with 10-month-olds, suggesting that these experimental results are consistent with a view of labels acting as features.

Plebe *et al.* (2010) [34] introduced a further word learning model consisting of parallel visual and auditory hierarchies. The authors used stages of training to simulate development, incorporating incrementally more layers of the hierarchy in learning. Only in their last, "linguistic," training phase are representations from the visual and auditory streams finally integrated. This stage-like process resulted in the emergence of "fast-mapping" at the time of integration.

In the following sections we introduce a new model to simulate the interaction between speech perception and object categorization. It draws on several of the above approaches in that it uses an interconnected pair of self-organizing maps, each of which represents a sensory domain (visual and auditory). In contrast to previous approaches, the focus here is on the interaction between the two maps *during learning*. Specifically, we introduce a novel learning mechanism that is suitable for integrating cross-modal information, in the sense that representations in one modality can be influenced by the other modality. As we shall see, this mechanism supports the emergence of categorical perception in both domains. Further simulations will demonstrate the dependence of developmental outcome on the timing of interactions, and also look at the effect of asymmetrical exposure to speech vs. visual objects, as is the case in infant development where speech is perceived prenatally, but objects occur only from birth onwards.

II. An Interactive Model of Word Learning

The model we propose combines two self-organizing maps [29], M_{vis} and M_{aud}, which are connected by Hebbian links ([28], [31], [32], [35], [36]). Fig. 1 presents a schematic illustration of the model's architecture. The maps M_{vis} and M_{aud} represent the visual and auditory domain respectively, and are trained with input from that domain. The two maps are further fully connected by bidirectional Hebbian links; i.e., every unit in the visual map, M_{vis}, is connected to every unit in the auditory map, M_{aud}. The learning algorithm involves updating the map weights (i.e., weights from the input to the map units) as well as the Hebbian connections. Importantly, the Hebbian weights

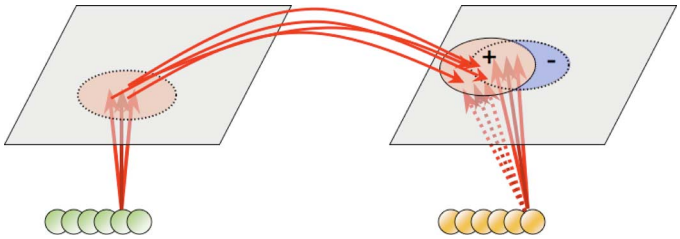

Fig. 1. Schematic illustration of the model architecture. Two self-organizing maps represent the auditory and visual domains. These are connected by Hebbian links which propagate activation between the maps and thus allow both modalities to influence each other. The elliptical regions and arrows illustrate in a simplified way the learning algorithm used for cross-modal interactions: direct activation is propagated via Hebbian connections to the other map, producing an indirect activation pattern. Weight update involves enhancement in the overlapping region of direct and indirect activation, and inhibition in units that are only activated by direct input.

are active during learning: this means that map-weight update in each modality is not just based on the activation resulting from an input pattern in the same modality, but also incorporates activation resulting from propagation of activation from the other modality. This way, auditory development can influence visual development, and vice versa.

Input patterns are presented to the model pairwise, i.e., one visual and one auditory pattern together. The presentation of an input pattern to each of the two maps generates a Gaussian activation pattern across each map. First, the best-matching units ($\mathrm{BMU}_{\mathrm{aud}}$ and $\mathrm{BMU}_{\mathrm{vis}}$) in the maps are calculated as the unit whose map vector is closest to the corresponding input vector, according to the distance $\mathcal{D}_{a,i}$ b etween the weight vector associated with unit i in M_a and input pattern $p = (p_1, \ldots, p_N)$ where $a \in \{vis, aud\}$ and $w_{k,i}$ the kth entry in the weight vector associated with unit i [see (1) and (2)]

$$\mathcal{D}_{a,i} = \sqrt{\sum_{k=1}^{N}(w_{k,i} - p_k)^2} \tag{1}$$

and

$$\mathrm{BMU}_a = \underset{i}{\operatorname{argmin}} \sqrt{\sum_{k=1}^{N}(w_{k,i} - p_k)^2}. \tag{2}$$

A Gaussian with standard deviation σ, centred on the BMU, is used as a coefficient for calculating unit activation. This activation pattern is termed the "direct activation" $\mathrm{act}_a^{\mathrm{dir}}$ [see (3)]

$$act_a^{\mathrm{dir}} = \mathcal{N}_{\mathrm{BMU}_a,\sigma}. \tag{3}$$

Activation is then propagated via the Hebbian links to the opposing map, resulting in an "indirect activation" pattern $\mathrm{act}_b^{\mathrm{ind}}$ (where $b \in \{aud, vis\}$ and $a \neq b$) on that map. This is a Gaussian over the maximally activated unit (MAU) after propagating activation though the Hebbian connections [see (4) and (5)]

$$\mathrm{MAU}_b = \underset{j}{\operatorname{argmax}} \sum_{i=1}^{N_a} act_{a,i} * h_{i,j}, \tag{4}$$

$$act_b^{\mathrm{ind}} = \mathcal{N}_{\mathrm{MAU}_b,\sigma} \tag{5}$$

where $a, b \in \{aud, vis\}$ and $a \neq b$, i the ith unit in map a, j the jth unit in map b, and $h_{i,j}$ the weight associated with the Hebbian connection between units i and j.

By propagating activation from the auditory to the visual map *and* from the visual to the auditory map, each individual map has a direct and an indirect activation pattern. From these, the joint activation $\mathrm{act}_a^{\mathrm{joint}}$ of map M_a, $a \in \{aud, vis\}$, is calculated for each unit i according to (6)

$$act_a^{\mathrm{joint}}(i) = (1 - \lambda) * act_a^{\mathrm{dir}}(i) + \lambda * act_a^{\mathrm{ind}}(i). \tag{6}$$

The parameter λ controls the impact of the cross-modal activation on learning, which can therefore change across development. Map update is performed by moving the weight vectors corresponding to active units (according to the joint activation patterns) closer to the current input pattern [see (7)]

$$w_{i,t+1}^a = w_{i,t}^a + \eta * act_a^{\mathrm{joint}}(i) * \mathcal{D}_{a,i} \tag{7}$$

where η is the learning rate.

However, as revealed by preliminary experiments, this cross-modal enhancement needs to be complemented by an inhibitory element. The idea behind the learning algorithm is that units receiving both direct and indirect activation represent a history of similar objects having been paired with similar labels (and vice versa). Visual map units receiving *only* direct activation may be activated by objects that are visually similar to objects previously activating this unit, but labeled differently, and therefore less likely to belong to the category corresponding to that unit. The map vector should therefore be moved away from the present input pattern. The corresponding logic holds for auditory units only receiving direct input. Therefore, in addition to the enhancing weight update, an inhibitory weight update is performed, resulting in the total weight update given in (8)

$$\begin{aligned} w_{i,t+1}^a = {} & w_{i,t}^a \\ & + \eta * act_a^{\mathrm{joint}}(i) * \mathcal{D}_{a,i} \\ & - \zeta * (act_a^{\mathrm{dir}}(i) - act_a^{\mathrm{joint}}(i)) * \mathcal{D}_{a,i}. \end{aligned} \tag{8}$$

After updating the map weights, Hebbian weights are strengthened for coactivated units (in the joint pattern) according to (10). Hebbian weights are further normalized to not grow larger than 1 [see(11)]

$$\Delta h_{i,j} = act_a^{\mathrm{joint}}(i) * act_b^{\mathrm{joint}}(j) \tag{9}$$

$$h_{i,j,t+1} = h_{i,j,t} + \kappa * \Delta h_{i,j} \tag{10}$$

$$h_{i,j,t+1}^{\mathrm{norm}} = \frac{h_{i,j,t+1} - \min(\mathcal{W})}{\max(\mathcal{W})} \tag{11}$$

where $\mathcal{W}$ is the set of all Hebbian weights.

Like Kohonen's (1982) algorithm, the present procedure incorporates several parameters. These are the neighborhood size σ and the learning rate η. In addition, this learning algorithm introduces the cross-modal integration coefficient λ, the inhibition coefficient ζ and the Hebbian update coefficient κ. While learning rate and neighborhood size decrease with time to enable learning first on a coarse, then on a finer scale, the other coefficients, which deal with cross-modal integration, increase over time. This reflects the fact that as organization in the individual maps becomes more reliable, this can be exploited more

TABLE I
PARAMETER SETTINGS FOR OPTIMAL LEARNING

Parameter	Initial setting	Change	Min/Max
η	0.1	$\eta_{t+1} = \eta_t * .99$	$\emptyset$
σ	3	$\sigma_{t+1} = \sigma_t^{-epoch/5000}$	$\emptyset$
κ	0.1	$\kappa_{t+1} = \kappa_t * 1.01$	0.3
λ	0.1	$\lambda_{t+1} = \lambda_t * 1.01$	1
ζ	0.001	$\zeta_{t+1} = \zeta_t * 1.015$	0.07

and more for cross-modal integration. The parameters are summarized in Table I.

A. Training Data

The visual stimuli were represented by geometrical surface-features measured from real objects. This kind of object representation has been used in other connectionist models of categorization (e.g., [37] and [38]), and similarity measures represented across these dimensions have been found to reflect infants' perception of the objects, as shown by their looking or object examination times [39]. Thus, toy objects from 11 categories (e.g., cat, squirrel, ship, car, table) were encoded using 18 different surface features. Each object was represented by an 18-dimensional vector of which each slot contained the (normalized) measured value of the corresponding surface feature. There were 8 objects from every category, resulting in a total of 88 object vectors. Word representations were kept as close to the acoustic wave form as possible, while also keeping the dimensionality of the resulting feature vectors low enough to make simulations computationally feasible. Therefore, a procedure similar to the one used by [32] was employed. In an infant's acoustic environment, every utterance ("token") of a word is different. Wave forms differ physically, depending on the speaker's voice, accent, intonation and syntactic context. Even several utterances from the same speaker differ in the length of individual segments and intonation. For this reason eight recordings of the 11 bisyllabic nonsense words (e.g., blicket, girru, sona, cruppet) were made, corresponding to the eight instances of every object category in the visual data set. The recordings were then preprocessed with Matlab. Each recording was sampled using 5 hamming windows. The short-time Fourier transform (STFT) was warped onto a mel frequency scale. Using the mel frequency scale for stimulus encoding means that the power spectrum of a sound is based on frequency bands equally spaced in the mel scale, which allows for a representation of the sound signal that approaches more closely that of human auditory perception. The discrete cosine transform (DCT) was then used to convert the frequency-domain signal back to the time domain. The result was a set consisting of 65-dimensional vectors representing the individual recordings. All of these preprocessing steps were performed using the RASTAMAT package for speech processing [40].

B. Model Evaluation

Model performance was assessed according to several criteria. On the one hand, the Hebbian connections encode the word-object mappings and are ultimately relevant for word learning performance – in other words, does hearing a word activate the appropriate visual representation (comprehension), and does presentation of a visual object result in the appropriate word (production)? On the other hand, as outlined above, we are particularly interested in the model's performance *within* the individual domains. Specifically, does category representation benefit from audiovisual interactions? For that reason, we assessed both mapping accuracy in both directions and the topographical organization in the maps. In particular, we used the following metrics, partially adapted from [41]. Most metrics we use here measure performance in terms of "projections" of a category; i.e., units that serve as best matching units for exemplars from the category.

1) Production: This measure corresponds to the capability of the model to produce the correct auditory label upon presentation of a visual signal. For every visual exemplar, its indirect projection via the Hebbian connections onto the auditory map was calculated, and the closest direct auditory projection was found. If this closest projection belonged to the auditory category that corresponded to the visual input, this exemplar's production score was set to 1. The overall map production score $Prod_M$ was then calculated as the proportion of exemplars that had production score 1.

2) Comprehension: This measure defines the model's capability of producing the correct visual image upon presentation of an auditory label, which is in turn related to word comprehension. The definition of the model's comprehension score $Comp_M$ was analogous to that of $Prod_M$.

3) Discrimination: This assesses the granularity of categorization, i.e., whether distinct exemplars are represented by distinct units or not. The discrimination capability of map M is defined as

$$\mathrm{Disc}_M = \frac{\sum_{c=1}^{C} \frac{\mathrm{Projections(c)}}{N}}{C} \tag{12}$$

where C is the number of categories, N the number of map units, and $\mathrm{Projections}(c)$ defines the number of unique units which are BMUs of exemplars of category c. In other words, the discrimination metric measures the average number of units onto which different exemplars from a category are projected. This value lies between 0 and 1, and small values correspond to low discrimination of exemplars, which forms part of the definition of categorical perception. Large values correspond to good discrimination.

4) Clustering: This metric measures the quality of categorization in the sense of good data clustering. For each exemplar of a category and its projection in the map, the neighborhood score NS evaluates the n (for n category members) nearest neighbor projections according to their category membership. If a neighboring projection also corresponds to the target exemplar's category, it contributes score $c = 1$, if it is from a different category, its score is $c = 0$. The neighborhood score NS is then defined as

$$NS = \frac{\sum_{i=1}^{n} c}{n}. \tag{13}$$

A *category's* neighborhood score is then the average neighborhood score of its members, $\mathcal{NS}_c = \sum_{i=1}^{n} NS/n$, and the whole model's clustering score is the average of all categories' neighborhood scores

$$\mathrm{Clust}_M = \frac{\sum_{c=1}^{C} \mathcal{NS}_c}{C}. \tag{14}$$

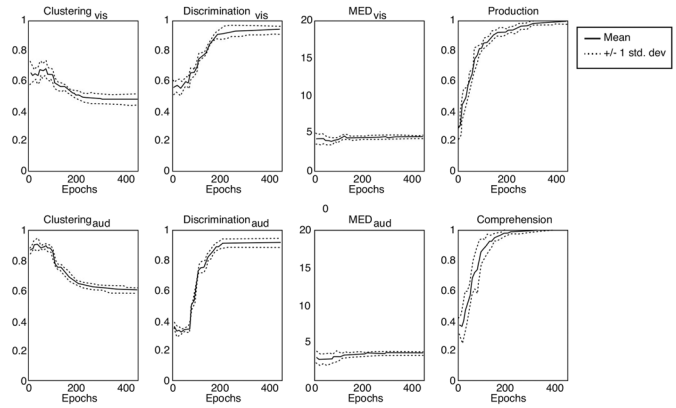

Fig. 2. Independent model performance: Hebbian weights were passive, so the maps could not influence each other during training. Top row: development in the visual map. Bottom row: Development in the auditory map.

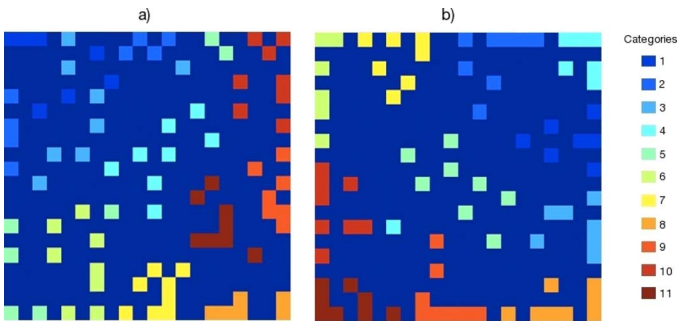

Fig. 3. Independent model: Exemplar projections after 450 epochs of training by category. a) visual map, b) acoustic map. a) Exemplar Projections (visual). b) Exemplar Projections (auditory).

5) Mean Exemplar Distance (MED): The MED metric refers to the averaged euclidean distance between the projections of two exemplars from the same category

$$\mathrm{MED}_M = \frac{\sum_{k=1}^{P} dist(p_k)}{P} \tag{15}$$

for all pairs p_k of exemplars from the same category, where $dist(p_k)$ is the euclidean distance between the two exemplars in pair p_k, P the number of pairs of exemplars in the category. Like $Clust_M$ this measures how well projections, i.e., BMUs, of category exemplars are grouped together. Instead of focussing on what categories *neighboring* units are activated by (in a way measuring how homogenous larger areas in map space are), this metric assesses whether projections form tight clusters in map space or are spread out.

III. Results

In order to investigate the impact of interactive learning, we trained two separate models – one using just the regular, noninteractive self-organizing map algorithm with "passive" Hebbian weights connecting the two maps (i.e., weight update within the maps is based solely on the direct activation pattern, merely the Hebbian connections encode coactivation of auditory and visual units), and one using the interactive algorithm presented above. Here, we report results with optimized parameter settings, which are given in Table I[1].

All results in the following two sections are based on 10 simulations with 11 categories. The map size chosen for this was a 20×20 grid of units. Since preliminary simulations showed that the maps settled into a stable state between 400 and 450 epochs of training with no changes occurring after this, all simulations reported here were carried out with a training phase of 450 epochs. One epoch consisted of a presentation of each pair of training exemplars (in random order), with weights being adjusted immediately.

A. Independent Maps

Fig. 2 shows the results of the model with independent map development. The model's behavior was similar with respect to Production and Comprehension scores. Initially both metrics increased rapidly, but started to level out once rates of approximately 85% of the mappings had been learned, after 130 (visual) and 120 (auditory) epochs, respectively. By 340 epochs, all mappings from the visual to the auditory domain had been encoded. Comprehension reached 100% first after 260 epochs. This comprehension/production asymmetry may stem from the smaller within-category variability in the auditory domain in this specific data set.

Discrimination in the visual domain reached its maximum, 93%, after 190 epochs of training[2], and then remained settled. There were thus 6 out of the 88 exemplars that were not mapped onto a separate unit in the map. Development in the auditory map was highly similar; the maximum discrimination of 91% was reached after 180 epochs.

The Mean Exemplar Distance metric was almost static at 4.4 in both the visual and auditory domains by 90 epochs, and only diverged from this value before the 90th epoch. This means that by this point in training, individual exemplars' projections were almost equally spaced out. This becomes clearer when examining Fig. 3. This plot depicts the 20×20 grid used for the simulations and the projections of exemplars after 450 epochs. The units activated by a given exemplar are highlighted in the color corresponding to its category. Exemplars are spread out evenly across the map in both domains.

B. Interactive Maps

Using the interactive algorithm, the model's behavior was radically different from that with independent maps (see Fig. 4). Development in the metrics reflecting Hebbian connection quality had a highly nonlinear time course. The Production measure exhibited a significant dip around 190–200 epochs. This was followed by an increase of production performance to 100% (where the metric settled at 320 epochs). The Comprehension metric, which was stable at 100% after 380 epochs, had a much less pronounced dip. This indicates that comprehension has an advantage above production early in training. Specifically, comprehension exceeded production between 100

[1]Optimal results were obtained after simulations with the following parameter variations: $\eta_{\min} = 0.01$, $\eta_{\max} = 0.5$, $\sigma_{\min} = 2$, $\sigma_{\max} = 8$, $\kappa_{\min} = 0.01$, $\kappa_{\max} = 0.3$, $\lambda_{\min} = 0.01$, $\lambda_{\max} = .3$, $\zeta_{\min} = 0.001$, $\zeta_{\max} = 0.1$

[2]During training, performance was measured after every 10th epoch. The maximum was taken to be achieved if the mean of 10 simulations after a fixed number of epochs was not significantly different from the maximum, as established via independent t-tests.

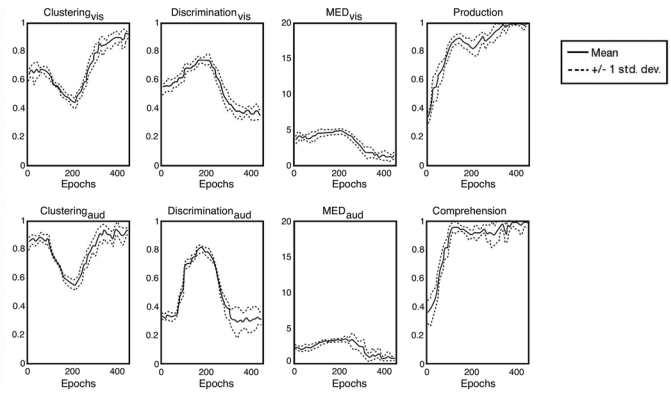

Fig. 4. Interactive model performance: the maps could influence each other through enhancement and inhibition. Top row: Development in the visual map. Bottom row: Development in the auditory map.

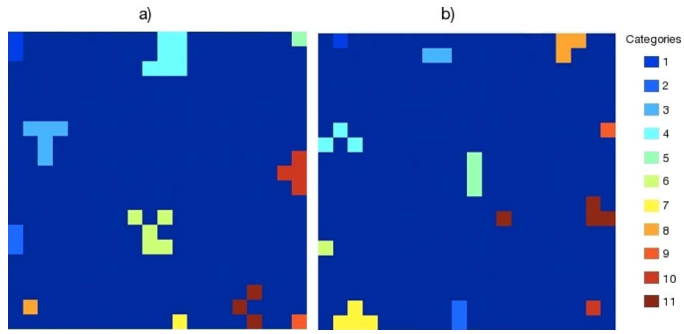

Fig. 5. Interactive model: Exemplar projections after 450 epochs of training by category. a) visual map, b) acoustic map. a) Exemplar Projections (visual). b) Exemplar Projections (auditory).

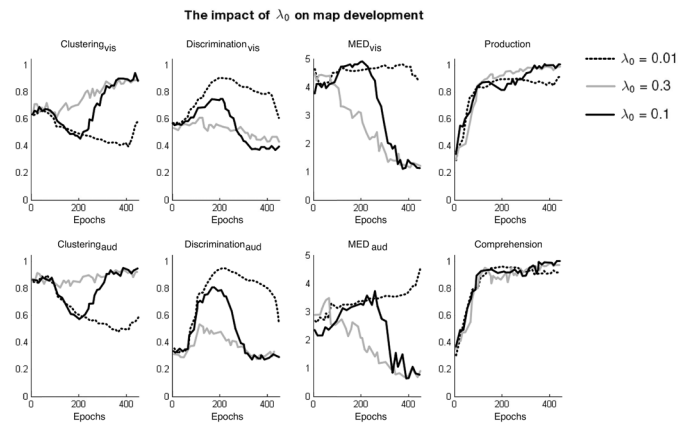

Fig. 6. Map development with different initial settings of the parameter λ, which defines the amount of map interaction: for values of $\lambda > 0.5$, *indirect* activation has a larger weight than *direct* activation as far as map update is concerned. Changing the initial setting of λ clearly has a dramatic impact on map development. Top row: Development in the visual map. Bottom row: Development in the auditory map.

and 210 epochs (based on two-tailed t-tests, significance level 0.05). This production-comprehension asymmetry probably arose from a higher within-category similarity of the auditory stimuli.

The clustering measure in the visual domain achieved a mean value of 90% during the last 30 epochs. This means that, on average, 90% of a projection's n nearest neighbor-projections (n being the number of exemplars in the category) were from the same category. Looking at the spatial distributions of projections in the map (see Fig. 5), it becomes clear why: The map has formed a representation of the categories in unit space whose behavioral consequence will be "categorical perception". Exemplars from one category are mapped onto tight clusters of units, whereas units representing different categories are far apart. This is the case in both the visual and auditory map, and is radically different from the map representations that emerged during independent map development, where within-category distances are about the same as between-category distances. Inspecting how map organization changed over time during training revealed an initial distributed representation of projections, similar to the independent model. However, between 170 and 200 epochs the maps started reorganizing, yielding tighter clusters of exemplar projections. This reorganization coincided with the dips in Comprehension and Production metrics and was also reflected in the three map evaluation metrics: discrimination in both domains increased for the first 170 epochs (180 in the auditory domain), but then dropped off steeply, settling at a value of 39% in the visual domain (i.e., approximately three patterns were mapped onto the same unit on average), and 30% in the auditory. The mean exemplar distance equally increased at first but then dropped to values around 1 (1.4 in the visual map, 0.74 in the auditory map). Clustering values decreased in the initial stages, but then began to increase from about 200 epochs onwards.

Closer inspection revealed that the reorganization phase began around the time in training at which λ was equal to 0.5, suggesting that the behavior is tied to the degree to which the maps influence each other. We therefore conducted a series of simulations varying the onset of λ in order to systematically examine the impact of early versus late interactions between the maps: keeping the growth of λ stable, a small onset value will yield later dominance of cross-modal activation, whereas a large onset value will lead to an early influence of – possibly immature – indirect activation patterns. These simulations are discussed in the next section.

IV. The Impact of Early vs. Late Map Interaction

In the settings used for the simulations presented above ($\lambda_0 = .1$) both maps were fairly well-developed by the time indirect activation played a significant role, even though cross-modal influence developed gradually from the start of training. Fig. 6 shows results for simulations with $\lambda_0 = 0.01$ and $\lambda_0 = 0.3$. In the case of $\lambda_0 = 0.01$, the point of "equal influence" is not reached until 380 epochs into training. As a consequence, map reorganization started late and was not completed by 450 epochs – a process which also appeared to have a detrimental impact on the quality of the Hebbian weights, as shown by the lower comprehension and production scores. Setting $\lambda_0 = 0.3$ by contrast led to a learning trajectory that apparently skipped reorganization: Clustering and MED appeared to increase/decrease almost steadily. The impact on comprehension and production was, however, fairly marginal.

The changes in developmental trajectories for different onset values of λ indicate that interactive development is not simple: the timing and amount of cross-modal influences is crucial. Even when the right learning mechanisms are in place, there is

no guaranteed advantage for interactive learning. Surprisingly, an earlier onset of cross-modal interactions ($\lambda_0 = 0.3$) did not lead to significant disruptions. It remains to be seen whether an even earlier audiovisual influence will lead to more errors in comprehension and production.

V. Asymmetrical Audio–Visual Development

While the onset of audiovisual interactions is clearly a relevant factor in the observed development of category and word representations as well as the mapping between them, these results raise another important question. The quality of the interactions are dependent on the structure in the individual domain maps – but this structure has its own, complex developmental trajectory in each sensory modality. While we have so far treated both auditory and visual processing as equivalent, this is not the case in reality. Besides differences that may be intrinsic to the nature of auditory vs. visual input from a signal-processing point-of-view, one very basic fact is that the two modalities do not commence development simultaneously. Previous research has demonstrated that infants' language development does not only start at birth: exposure to speech sounds begins in uterus. The sounds the foetus perceives are principally low-pass filtered maternal speech [42]. By contrast, exposure to visual stimuli from the environment is exclusively postnatal (as opposed to organism-internal retinal activation, which may *prepare* the visual system for environmental input, but not *stimulate* it in the same way, cf. [43]). Could this have an effect on the audiovisual interactions involved in word-object mappings? It appears likely that—in the SOM analogy—some structure is already present in the auditory map by the time true audiovisual events begin to occur. Although this is a coarse simplification, the impact of this type of asymmetry on an abstract level can be explored with simulations using the model presented above. To this end, we trained a map with the standard SOM algorithm on the auditory stimulus set for 70 epochs. This structured map then served as the starting point in the auditory domain for audiovisual training. Apart from this, no changes were made to the map architecture or parameter settings. Learning with the pretrained auditory map appeared to be highly similar to progress without pretraining. The only notable difference was an accelerated settling of the Comprehension score at 100%, which took 320 epochs–60 epochs less than in the standard training procedure. As the Hebbian weights, whose quality determines comprehension scores, were not involved (or modified) during pretraining, this represents an actual computational advantage. In order to investigate whether the *amount* of pretraining had an influence on cross-modal learning, the simulations were repeated with 200 epochs of auditory unimodal training before the onset of audiovisual events. This resulted in no noteworthy changes. Examining map representation at different stages during development showed the following: While the auditory map appeared well-structured at the onset of cross-modal training, the topographical organization acquired over the pretraining phase was lost within a few epochs of audiovisual exposure, and map-organization had to begin from scratch. The indirect activation being propagated into the auditory map from the (unstructured) visual map seemed to unravel the previously acquired category structure. A further set

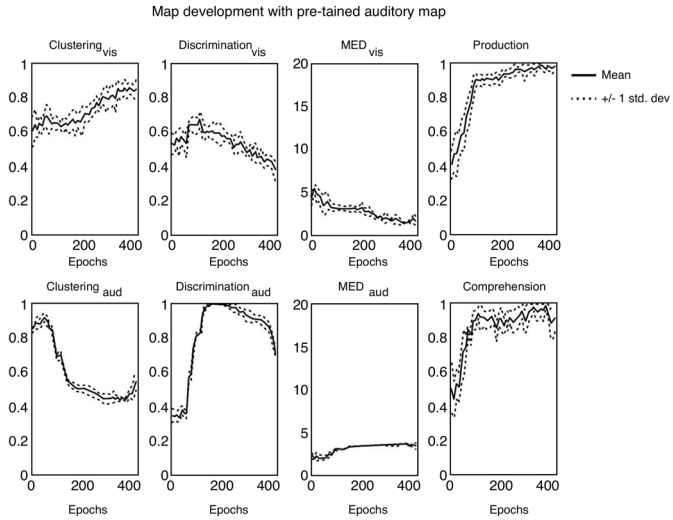

Fig. 7. Map development (interactive) after pretraining the auditory map for 70 epochs, with $\lambda_{\mathrm{vis}} = .6$ and $\lambda_{\mathrm{aud}} = .01$. Top row: Development in the visual map. Bottom row: Development in the auditory map.

of simulations therefore involved changing the initial parameter settings for cross-modal training. The target was to let the prestructured auditory map influence map development in the visual domain to a large degree (with the indirect auditory activation just overriding the direct activation), but to let the unstructured visual map have almost no impact on auditory development. Initial parameter settings were therefore $\lambda_{\mathrm{vis}} = 0.6$ and $\lambda_{\mathrm{aud}} = .01$ (all others remained the same, i.e., maps did not differ in plasticity). Fig. 7 shows the results of these simulations. The developmental trajectory clearly showed some changes, specifically in the visual domain. The "re-organization phase" seemed to be almost nonexistent in visual development, with the trajectories of Clustering and Mean Exemplar Distance, as well as Production, being almost monotonic. Similarly, the Discrimination metric only increased minimally before dropping off again, yielding a comparatively flat developmental trajectory. What this shows is that the pretrained map *can* have an impact on the opposite map, if the parameter settings are chosen accordingly. This is not necessarily an artificial construct: it seems plausible that the initial λ-setting can in some way be tied to the degree of map organization, and in terms of learning it definitely makes sense to prevent an unstructured map from altering a structured one, but to enhance the opposite. While the Comprehension and Production rate were not consistently significantly different from each other (independent t-tests), the model achieved a Comprehension rate of 100% after 330 epochs, and a Production rate of 100% after 440 epochs.

Even though the stimulus encodings in the present simulations were quite abstract and consequently the learning trajectories may not reflect the full complexity of processing speech versus visual objects, the simulation results suggest that this Comprehension/Production asymmetry may at least partially reflect the effect observed with infants, where word comprehension precedes production by several months. Asymmetric cross-modal learning may therefore play an underlying role in this observed discrepancy between infants' learning which objects words refer to, and actively using the words.

VI. General Discussion and Conclusion

We have presented a model of word learning that involves interacting self-organizing maps of visual and auditory representations. The learning mechanism this model introduces involves a "fine tuning" of learned representations that considers not just input in the corresponding domain, but also the "history" of similar input patterns having been paired with similar counterparts in the other domain, which is preserved in the Hebbian connections. As a result of this learning mechanism, categorical perception emerged in both domains: within-category distances between exemplars were small, but between-category distances large–yielding a qualitative change in category representations. This new category structure is computationally superior for treating stimuli categorically: assigning an object to a category is easy under these circumstances, as even marginal exemplars will be clearly mapped onto a category rather than in a transitional grey zone where similarity to members of two categories may prevent unambiguous category membership decisions. At the same time, smaller within-category differences can be ignored – meaning that features irrelevant for category assignment do not interfere with, e.g., selecting appropriate actions or expectations for encountered exemplars. This tightly clustered category representation developed with no cost in terms of word-object mapping accuracy. Although developmental trajectories in the interactive simulations exhibited a dip in mapping accuracy (Production and Comprehension scores), this was merely a transient effect as the auditory and visual domains began to reorganize with increased cross-modal influence. We hypothesize that this type of cross-modal interaction may play a role in infants' learning of word-object mappings. The initial effects of this mechanism may further underlie the reported facilitation of categorization through labeling ([11], [14], [16]), as the improved category structure exhibiting tighter clusters and large intercategory distances may enable infants to reject a test item as a member of the familiarized category—i.e., to exhibit increased novelty preference. The emerging category structure in the interactive model also suggests that discrimination between category members is decreased in comparison to unimodal (or independent) learning. This is an interesting hypothesis for future infant studies: None of the previous studies on infants' category formation in the presence of labels investigated specifically the encoding of familiarization items with regard to discrimination.

Further simulations revealed the role that the timing of cross-modal interactions plays in development. Early interactions appeared to improve learning, while a shift to dominance of cross-modal interactions only later in learning seemed to prevent the beneficial effects found earlier. Whether an even earlier onset of audiovisual interactions would be disruptive remains to be seen. However, it is possible to conclude from the simulations presented in this paper that it is not necessary—as has been suggested elsewhere [32]—to develop mature category representations before learning cross-modal connections. If anything, weak but early interactions seem more beneficial for a gradual learning than late interactions.

We finally looked at the role of asymmetric learning in this interactive scenario. While the prenatal onset of auditory learning compared to a postnatal onset of visual exposure can be disregarded in models where representations in both domains do not influence each other during training, it is crucial to ensure that development in any interactive model is robust enough to tolerate such asynchrony. Our simulations showed that this is the case—pretraining the auditory map did not have a negative impact on learning. Instead, with appropriate parameter settings this resulted in an advantage in word comprehension – an indicator that exposure asymmetry could be beneficial to audiovisual development in infancy.

Our model considers interactions between auditory and visual learning across development—essentially laying out a trajectory from the onset of auditory and visual input through to word-object mappings. While this is taking on a macro-perspective on a small scale (involving only a small number of visual and auditory categories) we think this is a valid illustration of how dynamic sensory representations (the two maps) interact over time and how small changes in each domain can give rise to complex developmental trajectories.

The present model only considers perceptual features of objects, rather than encoding *conceptual* attributes as well (e.g., "animate," [44]). This is not a general restriction – as a mathematical notion of similarity may be based on perceptual and conceptual features, the learning algorithms described here can apply to a much larger range of features. The general conclusion that learning about words and objects may occur in a parallel, interactive way – and such learning may in fact be advantageous – holds regardless of how the categories involved are defined.

The relationship between this model and existing studies of word learning and categorization is in a sense indirect as the time scale of the model is so different from infant experiments (usually involving "category learning" in a single session of about 1–2 min duration). The computational benefit of cross-modal learning explored here is, however, a plausible explanation for the advantage found in infant experiments for learning categories in the presence of labels. To construct a model simulating the experimental task directly is an aim for future work.

In conclusion, we have presented a plausible, cross-modally interactive model of word learning that can account for interactions between word processing and category formation in infancy. Future work will determine further whether the model is able to deal with larger scale lexical and object input, and also whether learning in the model could be made more biologically plausible by regulating internally some of the aspects currently handled by predefined parameters.

References

[1] P. Kuhl, "Early language acquisition: Cracking the speech code," *Nature Rev. Neurosci.*, vol. 5, pp. 831–843, 2004.
[2] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, pp. 1926–1928, 1996.
[3] H. Bortfeld, J. Morgan, R. Golinkoff, and K. Rathbun, "Mommy and me: Familiar names help launch babies into speech-stream segmentation," *Psychol. Sci.*, vol. 16, pp. 298–304, 2005.
[4] R. Tincoff and P. Jusczyk, "Some beginnings of word comprehension in 6-month-olds," *Psychol. Sci.*, vol. 10, pp. 172–175, 1999.
[5] W. Quine, *Word and Object*. Cambridge, MA, USA: MIT Press, 1960.
[6] P. Eimas and P. Quinn, "Studies on the formation of perceptually based basic-level categories in young infants," *Child Develop.*, vol. 65, pp. 903–917, 1994.
[7] B. Younger and L. Cohen, "Infant perception of correlations among attributes," *Child Develop.*, vol. 4, pp. 858–867, 1983.
[8] B. Younger, "The segregation of items into categories by ten-month-old infants," *Child Develop.*, vol. 6, pp. 1574–1583, 1985.

[9] B. Younger and L. Cohen, "Developmental change in infants' perception of correlations among attributes," *Child Develop.*, vol. 57, pp. 803–815, 1986.
[10] D. Mareschal and P. C. Quinn, "Categorization in infancy," *Trends Cogn. Sci.*, vol. 5, pp. 443–450, 2001.
[11] S. Waxman and D. Markow, "Words as invitations to form categories: Evidence from 12-to 13-month-old infants," *Cogn. Psychol.*, vol. 29, pp. 257–302, 1995.
[12] S. Waxman and I. Braun, "Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants," *Cognition*, vol. 95, pp. 59–68, 2005.
[13] M. Balaban and S. Waxman, "Do words facilitate object categorization in 9-month-old infants?," *J. Exp. Child Psychol.*, vol. 64, pp. 3–26, 1997.
[14] A. Fulkerson and S. Waxman, "Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds," *Cognition*, vol. 105, pp. 218–228, 2007.
[15] A. Ferry, S. Hespos, and S. Waxman, "Categorization in 3- and 4-month-old infants: An advantage of words over tones," *Child Develop.*, vol. 81, pp. 472–479, 2010.
[16] K. Plunkett, J. Hu, and L. Cohen, "Labels can override perceptual categories in early infancy," *Cognition*, vol. 106, pp. 665–681, 2008.
[17] C. Robinson and V. Sloutsky, "Linguistic labels and categorization in infancy: Do labels facilitate or hinder?," *Infancy*, vol. 11, pp. 233–253, 2007.
[18] N. Althaus and G. Westermann*, Labels Can Cause Infants to Split a Visual Category*, to be published.
[19] N. Althaus and D. Mareschal*, Labels Direct Infants' Attention to Commonalities during Novel Category Learning*, to be published.
[20] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cogn. Sci.*, vol. 26, pp. 113–146, 2002.
[21] C. Yu, "The emergence of links between lexical acquisition and object categorization: A computational study," *Connect. Sci.*, vol. 17, pp. 381–397, 2005.
[22] C. Yu, D. Ballard, and R. Aslin, "The role of embodied intention in early lexical acquisition," *Cognitive Sci.*, vol. 29, pp. 961–1005, 2005.
[23] P. Schyns, "A modular neural network model of concept acquisition," *Cogn. Sci.*, vol. 15, pp. 461–508, 1991.
[24] K. Plunkett, C. Sinha, M. Møller, and O. Strandsby, "Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net," *Connect. Sci.*, vol. 4, pp. 293–312, 1992.
[25] G. Schafer and D. Mareschal, "Modeling infant speech sound discrimination using simple associative networks," *Infancy*, vol. 2, pp. 7–28, 2001.
[26] C. Stager and J. Werker, "Infants listen for more phonetic detail in speech perception than in word-learning tasks," *Nature*, vol. 388, pp. 381–382, 1997.
[27] K. Yoshida, C. Fennell, D. Swingley, and J. Werker, "Fourteen-month-old infants learn similar-sounding words," *Develop. Sci.*, vol. 12, pp. 412–418, 2009.
[28] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Networks*, vol. 17, pp. 1345–1362, 2004.
[29] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, pp. 59–69, 1982.
[30] G. Carpenter and S. Grossberg, "Adaptive resonance theory," in *The Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. Cambridge, MA, USA: MIT Press, 2003.
[31] P. Li, X. Zhao, and B. MacWhinney, "Dynamic self-organization and early lexical development in children," *Cogn. Sci.*, vol. 31, pp. 581–612, 2007.
[32] J. Mayor and K. Plunkett, "A neurocomputational account of taxonomic responding and fast mapping in early word learning," *Psychol. Rev.*, vol. 117, pp. 1–31, 2010.
[33] V. Gliozzi, J. Mayor, J. Hu, and K. Plunkett, "Labels as features (not names) for infant categorization: A neurocomputational approach," *Cogn. Sci.*, vol. 33, pp. 709–738, 2009.
[34] A. Plebe, M. Mazzone, and V. de la Cruz, "First words learning: A cortical model," *Cogn. Comput.*, vol. 2, pp. 217–229, 2010.
[35] R. Miikkulainen, "Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon," *Brain Lang.*, vol. 59, pp. 334–366, 1997.
[36] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain Lang.*, vol. 89, pp. 393–400, 2004.
[37] D. Mareschal, R. French, and P. Quinn, "A connectionist account of asymmetric category learning in early infancy," *Develop. Psychol.*, vol. 36, pp. 635–645, 2000.
[38] R. M. French, D. Mareschal, M. Mermillod, and P. C. Quinn, "The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: Simulations and data," *J. Exp. Psychol.: General*, vol. 133, pp. 382–397, 2004.
[39] D. Mareschal, D. Powell, and A. Volein, "Basic-level category discriminations by 7-and 9-month-olds in an object examination task," *J. Exp. Child Psychol.*, vol. 86, pp. 87–107, 2003.
[40] D. P. W. Ellis, PLP and RASTA (and MFCC, and Inversion) in Matlab 2005 [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/
[41] F. Richardson and M. Thomas, "Critical periods and catastrophic interference effects in the development of self-organizing feature maps," *Develop. Sci.*, vol. 11, pp. 371–389, 2008.
[42] J.-P. Lecanuet, B. Gautheron, A. Locatelli, B. Schaal, A.-Y. Jacquet, and M.-C. Busnel, "What sounds reach fetuses: Biological and nonbiological modeling of the transmission of pure tones," *Develop. Psychobiol.*, vol. 33, pp. 203–219, 1998.
[43] R. Wong, "Retinal waves and visual system development," *Annu. Rev. Neurosci.*, vol. 22, pp. 29–47, 1999.
[44] J. Mandler, "Perceptual and conceptual processes in infancy," *J. Cogn. Develop.*, vol. 1, pp. 3–36, 2000.

Nadja Althaus received the Magister Artium degree in linguistics, psychology ,and computer science from the University of Tübingen, Tübingen, Germany. She received the Ph.D. degree in psychology from Birkbeck, University of London, London, U.K.

She currently holds the Winkler Career Development Fellowship in Experimental Psychology at St Hugh's College, University of Oxford, London, U.K. Her research focuses on interactions between visual and language learning in infants and combines computational models with infant eye tracking studies.

Denis Mareschal received the bachelor's degree in physics and theoretical physics from Cambridge University, Cambridge, U.K. He received the master's degree in psychology from McGill University, Montreal, QC, Canada, and then received the Ph.D. degree in psychology from Oxford University, London, U.K.

He is currently a Professor at Birkbeck University of London, London, U.K. His research centers on developing mechanistic models of perceptual and cognitive development in infancy and childhood.

Dr. Mareschal received the Marr Prize from the Cognitive Science Society (USA), the Young Investigator Award from the International Society on Infant Studies (USA), and the Margaret Donaldson Prize from the British Psychological Society. He is a Fellow of the Association for Psychological Science