Word Learning through Sensorimotor Child-Parent Interaction: A Feature Selection Approach

Chen Yu, Jun-Ming Xu* and Xiaojin Zhu* (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University
Bloomington, IN, 47405 USA
* Department of Computer Sciences, University of Wisconsin-Madison
Madison, WI, 53706 USA

Abstract

This paper presents a computational model of word learning with the goal to understand the mechanisms through which word learning is grounded in multimodal social interactions between young children and their parents. We designed and implemented a novel multimodal sensing environment consisting of two head-mounted mini cameras that are placed on both the child's and the parent's foreheads, motion tracking of head and hand movements and recording of caregiver's speech. Using this new sensing technology, we captured the dynamic visual information from both the learner's perspective and the parent's viewpoint while they were engaged in a free-play toy-naming interaction. We next implemented various data processing programs that automatically extracted visual, motion and speech features from raw sensory data. A probabilistic model was developed that can predict the child's learning results based on sensorimotor features extracted from child-parent interaction. More importantly, through the trained regression coefficients in the model, we discovered a set of perceptual and motor patterns that are informatively time-locked to words and their intended referents and predictive of word learning. Those patterns provide quantitative measures of the roles of various sensorimotor cues that may facilitate word learning, which sheds lights on understanding the underlying real-time learning mechanisms in child-parent social interactions.

Keywords: computational modeling, word learning, embodied cognition, perception and action.

Introduction

Just as in many other cognitive learning tasks, a critical problem in word learning is the uncertainty and ambiguity in the learning environment – young word learners need to discover correct word-referent mappings among many possible candidate words and many possible candidate referents from potentially many objects that are simultaneously available. At the macro level, we know a great deal about object name learning and how it seems to be characterized by attentional biases to attend to the shape of the whole object (Landau, Smith & Jones, 1998), by conceptual biases that make some kinds of word-meaning mappings more likely than others (Markman, 1989), and by all sorts of linguistic bootstraps whereby children use the words they already know to help figure out new meanings (Gleitman, 2005). But we know very little about how any of this works in real time and in the cluttered context of the real world interactions of toddlers and parents, contexts typically characterized by many interesting objects, many shifts in attention by each participant, and many goals (beyond teaching and learning words).

Previous studies have examined early word learning in constrained experimental tasks with only one or two objects in view. The adult partner (usually the experimenter) focuses on the child and on effective teaching, and provides clear and repeated signals of her attention to the object being named (E.g. Baldwin, 1993; Tomasello & Akhtar, 1995). In this way, the attentional task is simple, and easily described in discrete and categorical terms (the attended object vs. the distractor). These contexts are not at all like the real world in which word learning is embedded in *a stream of activity* -- in which parents both react to and attempt to control toddler behaviors and in which toddlers react to, direct, and sometimes ignore parents as they pursue their own goals.

To truly understand the mechanisms of word learning, we need to focus on more micro-level behaviors as they unfold in *real time* in the richly varying and dynamically complex interactions of children and their mature partners in more naturalistic tasks (such as toy play). Further, whereas the studies at the macro-level clearly demonstrate many intelligent behaviors in infant word learning, they have not yet led to a formal account of the underlying mechanisms. Thus, we want to know not only that learners use various cues in social interaction to facilitate learning (see a good example of macro-level modeling by Frank, Goodman & Tenenbaum, 2009), but also exactly *how* they do so in terms of the real-time processes in the naturalistic tasks wherein everyday language learning must take place.

To this end, we developed a novel paradigm with two critical components. First, we developed a multisensory experimental environment to capture multimodal data with the goal to study the dynamics of child-parent social interactions, that ultimately lead to word learning, at the sensorimotor levels – in the bodily gestures and as well as momentary visual and auditory perception of the participants. We developed various signal processing tools to automatically annotate such rich dataset. Second, we proposed and implemented a new probabilistic model based on state-of-the-art machine learning techniques to discover the perceptual and motor patterns that are informatively time-locked to words and their intended referents and *predictive* of word learning. In the following sections, we first describe our experimental setup and data. We then introduce our model of word learning. After that, we present the results from a set of simulation studies. Finally, we offer some general discussions and conclude our work.

Experiment

As shown in Figure 1, the naturalistic interaction of parents and toddlers in the task of table-top toy play was recorded

by three cameras from different perspectives: 1) A lightweight mini camera mounted on a sports-headband and placed low on the forehead of the child provided information about the scene from the *child learner's point of view*. This is a particularly important and novel component of our set-up. The angle of the camera is adjustable, and has a visual field of approximately 90°. 2) A ceiling camera provided a top-down third-person view, allowing a clear observation of exactly what was on the table at any given moment (mostly the participants' hands and the objects being played with). 3) Another head-mounted camera provided the parent's viewpoint. In addition, our multimodal system recorded participants' body movements through a motion tracking system as well as the parent's speech through a headset.

Participants. The target age period for this study was 18 to 20 months. We invited parents in the Bloomington, Indiana area to participate in the experiment. 13 dyads of parent and child were part of the study (5 male and 7 female). 7 additional children were not included because of failure to keep the head camera on. For the child participants included, the mean age was 19.6, ranging from 17 to 20 months. All participants were white and middle-class.

Stimuli. Parents were given two sets, with three toys in each set, in this free-play task. The toys were rigid plastic objects of simple shapes and were painted with one primary color. Each set had a red, a green and a blue object.

Procedure. The task was a common one in the everyday lives of children and parents – to take turns in jointly acting on, attending to, and naming objects. This is a common context in which children learn names for things. The toys used in this experiment were novel items. The child and parent sat opposite each other at a small table and the parent was instructed to interact naturally with the child, engaging their attention with the toys while teaching the words for them.

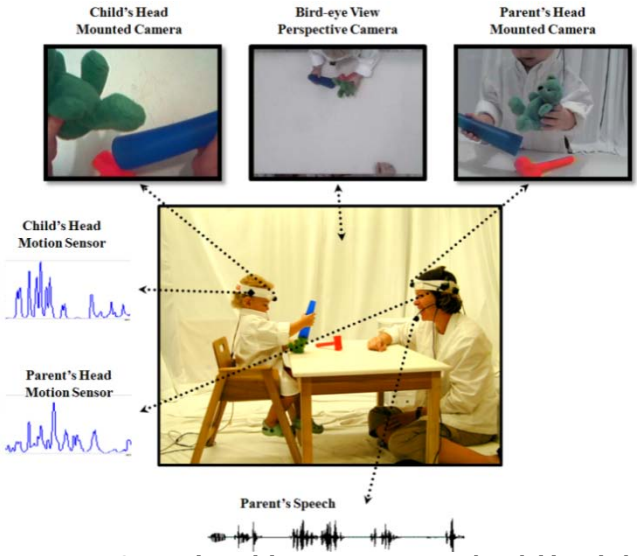


Figure 1. Our multimodal sensing system. The child and the mother played with a set of toys at a table. Two mini cameras were mounted on the child's and the parent's forehead, respectively to collect visual information from two first-person views. A third camera mounted on the top of the table recorded the bird's eye view of the whole interaction. They also wore motion sensors to track their head movements. A headset was used to record the parent's speech.

Parent-child free play session. The instructions given to the parent were to take all three objects from one set, place them on the table, play with the child and after hearing a command from the experimenters, remove the objects in this trial and move to the next set to start the next trial. Parents were given the names of the objects that they were to use and were instructed to teach the children those object names. However, there was no special instruction as to what the parents had to say or what they had to perform, just that they were to engage their child. All the names were artificial words. There were a total of four trials with each object set repeated twice, each about 1 minute long. The interaction between parent and child lasted between 4 and 7 minutes and was free-flowing in form.

Name-comprehension test. After the period of free interaction, the experimenter tested the child's comprehension of the object name for each of the 6 objects. This was done by placing three objects out of reach of the child about 30 inches apart, one to the left of the child, one in the middle, and one to the right. The experimenter then looked directly into the child's eyes, said the name of one of the objects and asked for it. For this portion of the experiment, a camera was focused on the child's eyes. Direction of eye gaze – looking to the named object when named – was scored as indicating comprehension. These recorded eye movements were coded (with the sound off) by a scorer naïve to the purpose of the experiment. Each word was tested twice with a score ranging from 0, 1 to 2: A word was given a score 2 if the child selected the correct target in both testing trials of that word, score 1 if the child successfully selected the correct one only once and score 0 if the child failed to select the correct one twice.

Multimodal Data and Data Processing

Given the multimodal data from child-parent interactions, we have developed various image and sensory processing tools to automatically annotate the data. This section briefly reviews our solutions to these problems. Technical details can be found in (Yu, Smith, Shen, Pereira, and Smith,2009).

Video Processing The recording rate for each of the three cameras was 30 frames per second. In the preliminary studies, there were 3 toy-play trials (with different sets of toys), each lasting about 60 seconds. Thus we collected approximately 24,300 (30 × 90 × 3 × 3) image frames from each interaction. The resolution of each image frame is 720*480. We analyzed the image data in two ways: (1) At the pixel level, we used the saliency map model developed by Itti, Koch, & Niebur (1998) to measure which areas in an image are most salient based on motion, intensity, orientation and color cues. Itti *et al.*'s saliency map model applies bottom-up attention mechanisms to encode for conspicuity (or ``saliency'') at every location in the visual input. (2) At the object level, the goal was to automatically

extract visual information, such as the locations and sizes of objects, hands, and faces, from sensory data in each of three cameras. These are based on computer vision techniques, and include three major steps. The combination of using pre-defined simple visual objects and utilizing start-of-the-art computer vision techniques resulted in high accuracy in visual data processing. The technical details can be found in (Yu, et al., 2009).

Motion data processing Six motion tracking sensors on participants' head and hands (3 sensors on each participant) recorded 6 DOF of their head and hand movements at the frequency of 240 Hz. Given the raw motion data from each sensor, the primary interest in the current work was the overall dynamics of body movements. We grouped the 6 DOF data vector into position {x, y, z} and orientation {h, p, r}. We then developed a motion detection program that computes the magnitudes of both position movements and orientation movements. In addition, we manually annotated which objects were in the child's or the parent's hands.

Speech processing We first segmented the continuous speech stream into multiple spoken utterances based on speech silence. Next, we asked human coders to listen to the recording and transcribe the speech segments. From the transcriptions, we calculated the statistics of linguistic information, such as the size of vocabulary and the average number of words per spoken utterance. Moreover, we extracted the onset and offset timestamps wherein an object name occurred in transcription and used them to define a naming event. In the next section, we will use these naming events to determine the learning patterns in visual and motion data streams.

As a result of our data processing, multiple heterogeneous time series were derived from multimodal raw data. In the present study, a set of 28 temporal sequences were selected which covered a wide range of sensorimotor dynamics in child-parent interaction, from the child's visual perception, to the child's hand and head actions, to the parent's hand and head actions. Figure 2 illustrates the meanings of some temporal variables.

The Model

We correlated the number of naming events for each object name with the score (0, 1 or 2) at test. We did not find a strong correlation between these two (r=0.1; p=0.28). The average number of naming events is 9 for a learned object name and 11 for an unlearned object name. Some object names that were provided just once or twice were actually learned, while others labeled by parents 5 or 6 times were not learned. This suggests that what matters were the specific contexts where those object names were named, what both parents and children visually attended to at those moments, and what they were doing at that time.

To discover those important sensorimotor features that led to successful learning through social interaction, we developed a formal model that predicts word learning results from multimodal features. Through the estimated

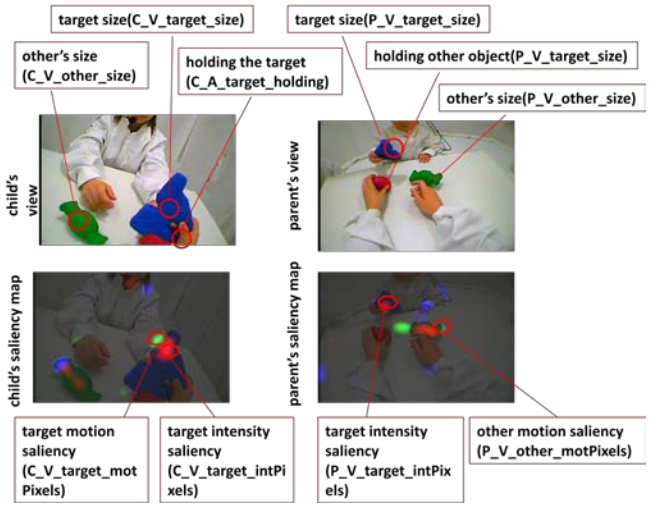


Figure 2. Derived temporal variables from multimodal data. Left: the child's view image and saliency map. Right: the parent's view image and saliency map. The naming of variables follows this standard: C(hild)/P(arent)_A(ction)/V(ision)_meaning. Eg. C_V_target_size means the size of the target object in the child's view.

regression coefficients, our main goal was to infer which features contribute to the learning outcome.

For notational simplicity, we begin by introducing our model in the case of a single parent-child interaction session, and only a single word was taught in that session. This will be generalized later. During the session the child is given $\#Events$ training naming events to learn the target word. These are encoded by some d-dimensional feature vectors $\mathbf{x}_1 \dots \mathbf{x}_{\#Events} \in \mathbb{R}^d$. As customary in machine learning, we assume that each feature vector is augmented by an extra dimension with constant value 1 for bias. We define the "gain" (from the child's perspective) from the k-th training naming event as

$$\mathbf{w}^\top \mathbf{x}_k$$

where $\mathbf{w}$ is a learning weight vector to be estimated. The total gain for the word is the sum over training naming events:

$$\sum_{k=1}^{\#Events} \mathbf{w}^\top \mathbf{x}_k = \mathbf{w}^\top \sum_{k=1}^{\#Events} \mathbf{x}_k \equiv \mathbf{w}^\top \mathbf{X}$$

where $\mathbf{X} \equiv \sum_{k=1}^{\#Events} \mathbf{x}_k$. Let $z \in \{0, 1\}$ be the hidden binary variable indicating whether the child actually learns the word. We model z with a logistic function:

$$P(z = 1 \mid \mathbf{w}, \mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{X})}$$

such that a larger total gain leads to a higher probability of learning. We cannot observe z directly. Instead, $\#Test$ test events are conducted after training. In each test event, the child had to choose the target object out of m different objects (m=3 in our case). Let $y_l \in \{0, 1\}$ be the observed variable on whether the child succeeded on the l-th test event, for $l = 1 \dots \#Test$. We assume that if the child has learned the word ($z = 1$), she would most likely pick the correct object (but there is a still probability that she may not pick the correct answer even when z = 1). This variability is captured by:

$$P(y_l \mid z=1) = \begin{cases} \gamma, & y_l = 1 \\ 1-\gamma, & y_l = 0 \end{cases}$$

where γ is a parameter less than 1. If the child does not learn the word ($z = 0$), we assume that she will randomly pick a test object with equal probability, resulting in

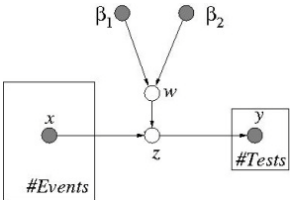
$$P(y_l \mid z=0) = \begin{cases} 1/m, & y_l = 1 \\ (m-1)/m, & y_l = 0 \end{cases}$$

These assumptions therefore model the likely noise in testing data.

To make a Bayesian probabilistic model, we introduce a prior distribution on $\mathbf{w}$. Sparse $\mathbf{w}$ (i.e., only a small number of features are relevant to learning) is preferred for interpreting the model, but we are also interested in including all related variables even they are pairwise correlated. Therefore, we employ the elastic net (Zou & Hastie, 2005), which corresponds to the prior,

$$P(\mathbf{w}) = h(\beta_1, \beta_2) \prod_{f=1}^{d} \exp\left(-\beta_1 |w_f| - \beta_2 w_f^2\right)$$

where β_1 and β_2 are non-negative parameters which control the tradeoff between prior and likelihood. The complete graphical model is given below.



We are now ready to state the modeling problem: Given training naming events $\mathbf{x}_1 \ldots \mathbf{x}_{\#\text{Events}}$ ($\mathbf{X} = \sum_{k=1}^{\#Events} \mathbf{x}_k$) and test outcomes $\mathbf{y} = (y_1 \ldots y_{\#Test})^\top$ (and hyper-parameters β_1 and β_2), what is the most likely weight coefficients $\mathbf{w}$? The non-zero (and large magnitude) elements can then be interpreted as the subset of features contributing to learning. The hidden variable z is of less interest, and is integrated out. Formally, we solve the *maximum a posteriori* (MAP) problem

$$\arg\max_{\mathbf{w}\in\mathbb{R}^d} \log P(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) = \arg\max_{\mathbf{w}\in\mathbb{R}^d} \log P(\mathbf{w}) \mathbf{p}(\mathbf{y} \mid \mathbf{w}, \mathbf{X})$$

The objective function can be equivalently written as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \beta_1 ||\mathbf{w}||_1 + \beta_2 ||\mathbf{w}||_2^2 + \log\left(\sum_{z=0,1} P(z \mid \mathbf{w}, \mathbf{X})\, P(\mathbf{y} \mid z)\right)$$

where $P(z \mid \mathbf{w}, \mathbf{X})$ and $P(\mathbf{y} \mid z)$ have been defined earlier.

Now, we are ready to extend the model to the multiple parent-child pairs and multiple words case. Let $i = 1 \ldots \#Pairs$ be the index for parent-child pairs. The i-th pair studied $j = 1 \ldots \#words(i)$ different words. Note the i-th pair's first word may be different than $i+1$-th pair's first word and so on. For the i-th pair's j-th word, there were $k = 1 \ldots \#Events(ij)$ training naming events. These naming events need not happen consecutively in time. We use $\mathbf{x}_{ijk} \in \mathbb{R}^d$ to denote the feature vector for the k-th naming event. Similarly, for the i-th pair's j-th word, there were $l = 1 \ldots \#Test(ij)$ test events y_{ijl}. In each test event, the child has to pick out the object corresponding to the word from m different objects.

We assume that all parent-child pairs share the same weight vector $\mathbf{w} \in \mathbb{R}^d$. For parent-child pair i on word j, the gain from learning experience k is

$$\mathbf{w}^\top \mathbf{x}_{ijk}$$

For parent-child pair i on word j the total gain from learning experiences is

$$\sum_{k=1}^{\#Events(ij)} \mathbf{w}^\top \mathbf{x}_{ijk}$$

In the same way, we can define z_{ij} and $P(y_{ijl} \mid z_{ij})$. The different pairs are independent to each other, so the *MAP* problem becomes

$$\min_{\mathbf{w}\in\mathbb{R}^d} \beta_1 ||\mathbf{w}||_1 + \beta_2 ||\mathbf{w}||_2^2 + \sum_{\mathbf{i}=1}^{\#\text{Pairs}} \sum_{\mathbf{j}=1}^{\#\text{Events(i)}} \log\left(\sum_{\mathbf{z_{ij}}=0,1} P(z_{ij} \mid \mathbf{w}, \mathbf{X}_{ij}) P(\mathbf{y_{ij}} \mid z_{ij})\right)$$

This optimization problem is non-convex and we optimize it with the Constrained Concave Convex Procedure (CCCP) (Yuille & Rangarajan, 2003).

Simulations and Results

Several parameters have to be set before we can identify interesting features with our model. First, in our model, we assume that the child may choose an incorrect object in the test events even when $\boldsymbol{z} = \mathbf{1}$ and denote the probability as $\boldsymbol{\gamma}$. However, there is no well-studied $\boldsymbol{\gamma}$ we can use. $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the weights of regularizers, which control the tradeoff between fitness to data and model complexity. To set them appropriately, we first chose several candidate parameters and use cross-validation to choose the best settings based on log-likelihood on the training set. Intuitively, $\boldsymbol{\gamma}$ should be greater than 0.5 and therefore the candidate set of $\{\mathbf{0.6}, \mathbf{0.7}, \ldots, \mathbf{1.0}\}$ was chosen. We tested several different $\boldsymbol{\beta_1}$ values $\{\mathbf{10^{-4}}, \mathbf{10^{-3}}, \ldots, \mathbf{10^{2}}\}$ to produce different solutions with different levels of sparsity and meanwhile the candidates of $\boldsymbol{\beta_2}$ were evaluated with a larger but sparser grid $\{\mathbf{10^{-6}}, \mathbf{10^{-4}}, \ldots, \mathbf{10^{4}}\}$. The thirteen parent-child pairs were randomly split into seven folds. Each time, one fold was left out as a tuning set and a model was trained on the remaining folds with each combination of three candidate parameters. The parameter setting with the highest average tuning set log-likelihood was selected, which is $\{\boldsymbol{\gamma} = \mathbf{0.7}, \boldsymbol{\beta_1} = \mathbf{0.001}, \boldsymbol{\beta_2} = \mathbf{0.01}\}$. We fixed this setting in the following experiments.

Through applying the model to sensorimotor features and showing that the model can predict word learning results ($\mathbf{y}$=0, 1 or 2) based on cross-validation, our main goal of the present study here was to gather and analyze the weights $\mathbf{w}$ of sensorimotor features $\mathbf{x}$ from training, and interpret those weight results to better understand what sensorimotor features may be predictive to learning results and therefore contribute to successful learning. To this end, our first study examined sensorimotor dynamics around object naming events with the assumption that what happened at those moments was more relevant to learning than other moments. In practice, we decomposed the whole training dataset into three groups based on when object naming events happened:

1) "during" moments defined by the onset and offset of a naming event; 2) "before" moments defined by 5 seconds prior to the onset of a naming event to that onset; 3) "after" moments defined by the offset of a naming event to 5 seconds after that offset. For each group based on the above timing definition, we extracted 28 features from the continuous time series (a subset of them were illustrated in Figure 2) and fed the data into the model as feature vector **x** to predict **y**. This is done separately for before, during and after moments.

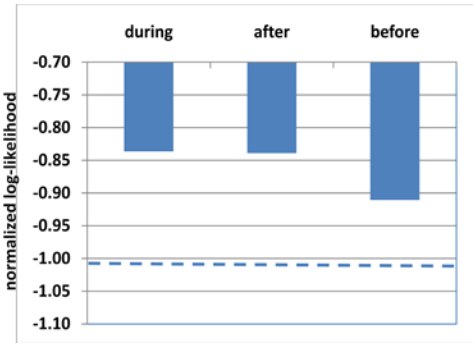


Figure 3. Normalized log-likelihood values from three temporal groups compared with the baseline (the dotted line).

First, the log-likelihood results shown in Figure 3 indicate the fitness of data from those three moments compared with a baseline calculation. Sensorimotor features from "during" and "after" moments are more predictive to word learning than those features from "before", while all three features sets are predictive as compared with chance (the dotted line in Figure 3). Next, we asked exactly what features in each moment are predictive to learning. This information can be inferred from the trained regression coefficients **w**. In practice, we selected top 5 features that have gained largest absolute weights based on training (note weights can be positive or negative). As shown in Figure 4, some sensorimotor features consistently played a role in all of the three temporal moments. For example, the child's holding of the target object (C_A_target_holding) appeared to be a predictive cue in all of the three moments, suggesting that the target object held by the learner is more likely to be learned. Manually holding the named object around the moments of hearing that object name indicates the learner's sustained attention and interest on the named object. In addition, in both "during" and "after" moments, the size of the target object in the child's view (C_V_target_size) and the parent's holding of the target object (P_V_target_holding) are good for learning but probably for different reasons. The size of target object in the child's view is a direct measure of visual saliency of that object – an indicator of the learner's attention. On the other hand, holding the target object by parents may facilitate learning if and only if this action can attract the child's attention – an open question worth more studies. Moreover, the stability of the child's head (C_A_head_rotSpeed) before and during naming also predicts good learning as compared with other cues. That is, the learner not only paid attention to the right object, manually held the object, but also stabilized their

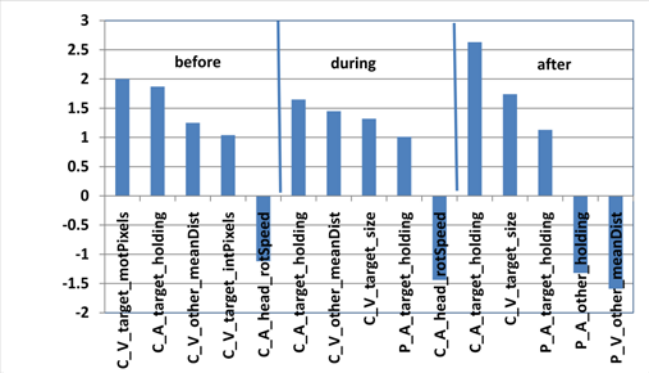


Figure 4. The top 5 features within before/during/after naming events that are critical for predicting word learning results.

attention for a certain period of time. All these led to better learning.

Since there were different types of sensorimotor features in training data, our next study focused on discovering which types of features are more important. To do so, we divided 28 features into four semantic subsets: child's perception, parent's perception, hand action and head movement. We then fed the features in each subset into the model and asked the model to predict word learning only based on those features within each subset. Figure 5 shows the results of log-likelihood values, suggesting the child's visual perception is more directly predictive than both the child's and the parent's actions. One plausible explanation is that the ultimate role of actions in learning is to select visual information for the internal learning processes. Interestingly, just head movements of the child and the parent can also somehow predict learning (not perfectly but far above chance). The stability and dynamics of head movements are good indicators of sustained attention of the learner and the teacher if they jointly attend to the same object. Lastly, the parent's perception is less relevant to learning, suggesting that the parent's perception may determine what action the parent may generate next and this next action can indirectly influence the child's action and the child's perception, but the parent's perception may not be directly relevant to learning.

We next closely examined each group and analyzed what features in each group contribute more toward predicting word learning results. Due the page limit, we selected only two most influential groups (child's perception and hand actions) and within each group, we selected only two most influential features to show in Figure 6. The child's holding of the target object didn't matter before/after naming moments but played a critical role exactly during naming moments. In contrary, the parent's holding of the target object had a negative weight before naming, no influence during naming, and became critical after naming. There are two plausible interpretations of those patterns. One is that the child held the target object while the parent named it and then passed that object to the parent's hands. The other possibility is that those patterns were mixed from two different interaction modes, both of which can lead to

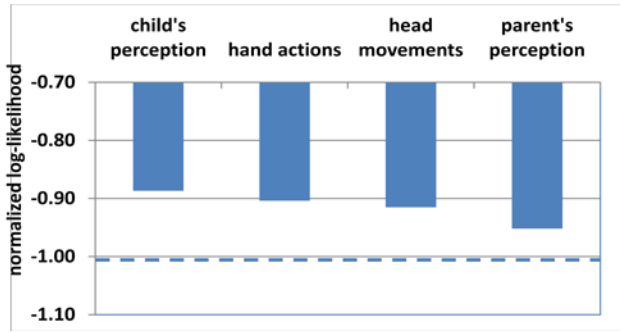


Figure 5. Normalized log-likelihood values of the four semantic feature groups.

successful learning. In one mode, the child led the interaction by holding the named object during naming. In the other mode, the parent named the object first and manually held the object to attract the child's attention after naming it. We need further studies to understand this better. Also shown in Figure 6 (right), the sizes of objects (both the target and other objects) seem to be a more direct measure of what the child visually attended to compared with other visual features (e.g. distance to the center or visual saliency). In particular, the size of the target object in the child's view is weakly relevant before and after naming, but right at the naming moment, this visual property seems to be critical. In contrast, the size of other objects played a negative role after naming. It is an open question why the size of other objects has a positive impact during naming.

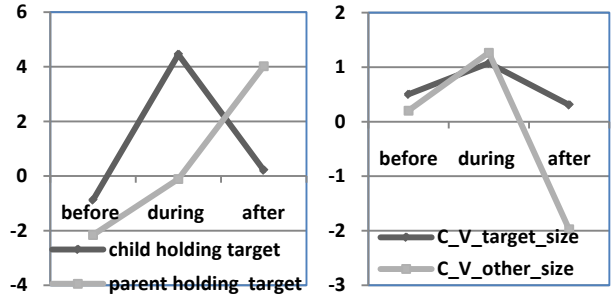


Figure 6. Left: the weights of the child's holding and the parent's holding of the target object. Right: the weights of the target object size and the size of other objects.

General Discussions and Conclusion

Most of children's word learning takes place in messy contexts – like the tabletop play task used here. There are multiple objects, multiple shifts in attention and multiple bodily cues by both partners, and many object names simultaneously available. In this paper, we used advanced sensing equipment and state-of-the-art experimental paradigms to collect multiple streams of real-time sensory data in parent-child interactions. Given such fine-grained data, we developed a formal model to analyze these multisensory data and to extract statistical regularities in the physical and social learning environment. We conducted two simulation studies to address questions such as which types of sensorimotor features are more important for learning, what moments are the right moment for the teacher to name objects and for the learner to build a mapping between names and objects, and how those feature may work together to facilitate learning. Those results derived from our modeling efforts (e.g. the regression coefficients of features) provided quantitative measures of how various bodily cues and sensory features may be relevant to learning at different moments and through different ways. Some results confirmed our original hypotheses and others were rather surprising, opening up new research questions with the potential to lead to new findings that we do not know yet. The present paper represents our first efforts in modeling sensorimotor dynamics in child-parent social interaction. With more fine-grained data and advanced computational modeling methods, we have the opportunity to discover a more complete mechanistic explanation of early word learning.

Acknowledgments: We thank Charlotte Wozniak, Amanda Favata, Alfredo Pereira, Amara Stuehling, and Andrew Filipowicz for data collection, and Thomas Smith for developing data management and preprocessing software. This research was supported by NSF BCS 0924248, NSF IIS-0953219, AFOSR FA9550-09-1-0665 and AFOSR FA9550-09-1-0313.

References

Baldwin, D. (1993). Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental psychology*, *29*, 832-843.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. (2008). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems, 20.*

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*(1), 23-64.

Itti, L., Koch, C., Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **20**, No. 11, pp. 1254-1259.

Landau, B., Smith, L. B., & Jones, S. S. (1998). Object shape, object function, and object name. Journal of Memory & Language, 38 (1), 1–27.

Markman, E. M. (1989). Categorization and naming in children. Cambridge, MA: MIT Press.

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*, 201-224.

Yu, C., Smith, L., Shen, H., Pereira, A., & Smith, T. (2009). Active Information Selection: Visual Attention Through the Hands. *IEEE Transactions on Autonomous Mental Development*, *2*, 141–151.

Yuille, A.L. & Rangarajan A. (2003). The concave-convex procedure. *Neural Computation*,15(4):915–936.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.