



ELSEVIER

ScienceDirect

Neurocomputing ■ (■■■■) ■■■-■■■

NEUROCOMPUTING

www.elsevier.com/locate/neucom

A unified model of early word learning: Integrating statistical and social cues

Chen Yu^{a,*}, Dana H. Ballard^b^a*Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University, Bloomington, IN 47405, USA*^b*Department of Computer Science, University of Rochester, Rochester, NY 14627, USA*

Received 30 January 2005; received in revised form 3 January 2006; accepted 30 January 2006

Abstract

Previous studies on early language acquisition have shown that word meanings can be acquired by an associative procedure that maps perceptual experience onto linguistic labels based on cross-situational observation. Recently, a social-pragmatic account [M. Tomasello, Perceiving intentions and learning words in the second year of life, in: M. Bowerman, S. Levinson (Eds.), *Language Acquisition and Conceptual Development*, Cambridge University Press, Cambridge, 2000, pp. 111–128] focuses on the effect of the child's social-cognitive capacities, such as joint attention and intention reading. This paper argues that statistical and social cues can be seamlessly integrated to facilitate early word learning. To support this idea, we first introduce a statistical learning mechanism that provides a formal account of cross-situational observation. A unified model is then presented that is able to make use of different kinds of embodied social cues, such as joint attention and prosody in maternal speech, in the statistical learning framework. In a computational analysis of infant data, our unified model performs significantly better than the purely statistical approach in computing word–meaning associations.

© 2007 Published by Elsevier B.V.

Keywords: Word learning; Language development; Computational modeling; Cognitive development; Social cue; Statistical learning

1. Introduction

Quine in [35] presented the following puzzle to theoreticians of language learning: imagine that you are a stranger in a strange land with no knowledge of the language or customs. A native says “gavagai” while pointing at a rabbit off in the distance. How can you determine the intended referent? Quine offered this puzzle as an example of *reference uncertainty*. Given any word–event pairing, there are, in fact, an infinite number of possible intended meanings—ranging from the rabbit as a whole, to its color, fur, parts, or activity. You need to find what “gavagai” exactly refers to. In real world, young children face with a harder problem in first language acquisition compared with Quine's example: (1) they hear continuous speech consisting of multiple words instead of a

single isolated word and (2) the parents do not always point to the location to narrow the range of relevant perceptual information and in doing so constrain the range of intended meanings. Thus, there are multiple words on the language side and multiple possible referents on the meaning side. Young language learners need to figure out which word goes to which meaning from multiple temporally cooccurring word–referent pairs.

A common conjecture of word learning is that children map sounds to meanings by seeing an object while hearing an auditory word form. The most popular mechanism of this word-learning process is *associationism*, which concentrates on statistical learning of cooccurring data from speech and extralinguistic context (see a review by Plunkett [33]). Richards and Goldfarb [39] proposed that children come to know the meaning of a word through repeatedly associating the verbal label with their experience at the time that the label is used. Smith [45] argued that word learning is initially a process in which children's attention is captured by objects or actions that are the most salient in

*Corresponding author. Tel.: +1 812 856 0838.

E-mail addresses: chenyu@indiana.edu (C. Yu), dana@cs.rochester.edu (D.H. Ballard).

1 their environment, and then they associate it with some
2 acoustic pattern spoken by an adult. The associative
3 approach, however, has been criticized on the grounds
4 that it does not provide a clear explanation about how
5 infants map a word to a potential infinity of referents when
6 the word is heard—the *reference uncertainty* problem
7 pointed out by Quine [35].

8 In the present article, we attempt to provide computa-
9 tional accounts of how young children build word-to-world
10 mappings. As an infinite number of meanings are
11 presented, they may use some kinds of constraints to guide
12 word learning by reducing the search space. This article
13 explores the role of two types of constraints: (1) statistical
14 regularities of the cooccurrences of words and meanings in
15 cross-situational observation and (2) social cues encoded in
16 multimodal parent–child interaction. We are not the first to
17 suggest that children may utilize those constraints. None-
18 theless, the major contribution of this work is to present a
19 detailed computational mechanism to show how multiple
20 cues may converge to provide reliable links between words
21 and the world. More specifically, we present a unified
22 model wherein statistical and social cues are integrated in a
23 single general framework. The simulation studies in this
24 paper provide a quantitative analysis of the role of these
25 cues in early word learning.

26 The organization of the paper is as follows: we first
27 review two accounts of word learning—a statistical
28 account and a social-pragmatic account. Then Section 3
29 proposes our unified model that integrates statistical and
30 social cues in a general system. Section 4 describes both the
31 data used in this work and preliminary statistical analyses
32 of the data. Section 5 presents the implementation of a
33 statistical learning model, which provides a probabilistic
34 framework using multiple word–meaning pairs collected
35 across different learning situations to compute distribu-
36 tional statistics and then establish word-to-world map-
37 pings. Section 6 describes the methods to extract prosodic
38 cues from raw speech and joint-attention cues from
39 infant–caregiver interaction. An integrative mechanism
40 based on the unified model utilizes social cues as spotlights
41 to highlight both words in speech and speakers’ intended
42 referents in the physical environment. Section 7 provides a
43 comparative study of different methods considering
44 different sets of statistical and social cues.

45 2. Related work

46 This section reviews two well-known accounts of
47 language acquisition. A statistical learning view suggests
48 that language acquisition is a statistically driven process in
49 which young language learners utilize the lexical content
50 and syntactic structure of speech as well as extralinguistic
51 contextual information as input to compute distributional
52 statistics. The social-pragmatic account focuses on “mind
53 reading” (social cognition) as fundamental to the word-
54 learning process. Both accounts have been supported by
55 various empirical and computational studies.

2.1. A statistical account

59 One explanation of how infants discover one-to-one
60 correspondences between multiple spoken words and their
61 meanings, termed “cross-situational learning”, has been
62 proposed by many theorists, such as Pinker [32] and
63 Gleitman [16]. This scheme suggests that when a child
64 hears a word, she can hypothesize a set of the potential
65 meanings for that word from the non-linguistic context of
66 the utterance containing that word. Upon hearing that
67 word in several different utterances, each of which is in a
68 different context, she can intersect the corresponding sets
69 to find those meanings which are consistent across the
70 different occurrences of that word. Presumably, hearing
71 words in enough different situations would enable the child
72 to rule out all incorrect hypotheses and uniquely determine
73 word meanings. However, there are few studies that
74 provide quantitative evidence of this mechanism.

75 A new trend in language acquisition focuses on the
76 importance of distributional information (e.g. [43,30,33];
77 see a review in [41]). The claim is that human language
78 learners including adults, children, and even infants possess
79 powerful statistical learning capacities. It is encouraging
80 that recent experimental evidence demonstrates that the
81 cognitive system is highly sensitive to distributional
82 features of the input (e.g. occurrence statistics). Among
83 others, Saffran et al. [43] showed that 8-month-old infants
84 are able to find word boundaries in an artificial language
85 only based on statistical regularities in speech. Later
86 studies [42] demonstrated that infants are also sensitive to
87 transitional probabilities over tone sequences, suggesting
88 that this statistical learning mechanism is more general
89 than the one dedicated solely to process linguistic data.
90 More recently, Newport and Aslin [30] showed that human
91 learners are able to discover non-adjacent distributional
92 regularities among speech sounds. Furthermore, Maratsos
93 and Chalkley [27] suggested that grammatical categories
94 could also be learned through a distributional analysis of
95 the speech input. Recent analyses on child-directed corpus
96 (e.g. [36,29]) demonstrated that simple computational
97 mechanisms using distributional information as a powerful
98 cue can obtain a considerable amount of knowledge on
99 grammatical category membership.

101 2.2. A social-pragmatic account

102 A major advance in recent developmental research has
103 been the documentation of the powerful role of social-
104 interactional cues in guiding infants’ learning and in
105 linking the linguistic stream to objects and events in the
106 world [1,50]. The social-pragmatic account of language
107 acquisition argued that the major sources of constraints in
108 language acquisition are social-cognitive skills, such as
109 children’s ability to infer the intentions of adults as adults
110 act and speak to them [1,49,7]. These kinds of social
111 cognition are called “mind reading” by Baron-Cohen [4].
112 Kuhl et al. [23] studied whether phonetic learning of 9–10

1 month infants is simply triggered by hearing language. If
 2 so, children should be able to learn by being exposed to
 3 language materials via digital video without human
 4 interaction. However, the results showed that infants could
 5 not learn phonetics through this way, suggesting that the
 6 presence of a live person provides not only social cues but
 7 also referential information. Butterworth [9] showed that
 8 even by 6 months of age, infants demonstrated sensitivities
 9 to social cues, such as monitoring and following another
 10 person's gaze. In Baldwin's work [1], the 18-month-old
 11 infant heard the novel word while his/her attention was
 12 focused on one toy and the experimenter looked at another
 13 toy. When children heard the same word in a testing phase,
 14 they chose the object at which the experimenter had been
 15 looking. This suggested that the infants were able to follow
 16 the speaker's attention and infer the mental state of the
 17 speaker to determine the referent of the novel word.
 18 Similarly, Tomasello [49] showed that infants were able to
 19 determine adults' referential intentions in complex inter-
 20 active situations. In one study, the novel word was not
 21 uttered during the time that the object was presented.
 22 However, infants could still build the correct word-referent
 23 association by retaining the label and waiting for the
 24 introduction of the object. In another study, children were
 25 able to connect the word with the action the speaker
 26 seemed satisfied with but not the one that was close in
 27 temporal proximity. Thus, they made use of the experi-
 28 menter's social cues to determine which action or verb was
 29 the intended referent. Tomasello concluded that the
 30 understanding of intentions, as a key social-cognitive skill,
 31 is the very foundation on which language acquisition is
 32 built.

33 2.3. Modeling word learning

34 A computational model can be used as an ideal observer
 35 to demonstrate whether a simulated learning device is able
 36 to induce statistical patterns when it is fed with the data
 37 similar to those that human learners perceive. In this way,
 38 simulation studies can provide not only useful hints but
 39 also insightful predictions for further experimental studies.

40 Several modeling approaches have capitalized on statisti-
 41 cal or logic algorithms that can learn cross-situational
 42 high-probability cooccurrences. MacWhinney [26] applied
 43 the competition theory to build an associative network that
 44 was configured to learn which word among all possible
 45 candidates referred to a particular object. Siskind [44]
 46 developed a mathematical model based on cross-situational
 47 learning and the principle of contrast, which learned
 48 word-meaning associations when presented with paired
 49 sequences of presegmented tokens and semantic represen-
 50 tations. Tenenbaum and Xu [48] developed a computa-
 51 tional model based on Bayesian inference which could infer
 52 meanings from one or a few examples without encoding the
 53 constraint of mutual exclusion. Roy and Pentland [40] used
 54 the correlation of speech and vision to associate spoken
 55 utterances with a corresponding object's visual appearance.

56 The learning algorithm was based on cross-modal mutual
 57 information to discover words and their visual associa-
 58 tions.

59 Several other approaches have accounted for different
 60 aspects of behavioral observations in child language
 61 development. Plunkett et al. [34] built a connectionist
 62 model of word learning in which a process termed
 63 autoassociation mapped preprocessed images with linguis-
 64 tic labels. The linguistic behavior of the network exhibited
 65 non-linear vocabulary growth (vocabulary spurt) that was
 66 similar to the pattern observed in young children. Colunga
 67 and Smith [10] presented an approach that extracted the
 68 correlations characteristic of the first 300 nouns that
 69 children learned. The results showed that regularities
 70 among object and substance categories were learnable
 71 and generalizable, enabling the system to become, after
 72 training, a more rapid learner of new object and substance
 73 names. Regier's work focused on grounding lexical items
 74 that described spatial relations in visual perception [37]. Li
 75 et al. [25] proposed a self-organizing-map-based develop-
 76 mental model that learned topographically organized
 77 representations for linguistic categories over time. In Yu
 78 et al. [52], egocentric multisensory data were used to first
 79 spot words from continuous speech and then associate
 80 action verbs and object names with their perceptually
 81 grounded meanings. The central idea was to utilize body
 82 movements as deictic references to associate temporally
 83 cooccurring data from different modalities.

84 The model in the present paper is different from previous
 85 work in that (1) we suggest that social cues in multimodal
 86 mother-child interaction can be embodied by the mother's
 87 animated actions, such as eye gaze direction, body
 88 orientation, and prosody in speech; (2) we argue that
 89 grounding social cues at the sensorimotor level leads to the
 90 integration of social and statistical cues in a single general
 91 system; and (3) we test our model using realistic data
 92 collected from mother-child interaction.

93 3. A unified model

94 Bloom [6] argued that children's conceptual biases,
 95 intentional understanding, and syntactic knowledge are
 96 not only necessary for word learning but that they are also
 97 sufficient. This claim contrasts with the theory that a
 98 fundamental mechanism of word learning is based on an
 99 associative process sensitive to statistical properties (cooc-
 100 currence of words and referents, etc.) of the input [33]. The
 101 associative view suggests that the child's sensitivity to
 102 spatio-temporal contiguity is sufficient for word learning,
 103 as postulated by associationist models of language
 104 acquisition with support by computational implementation
 105 [14,38]. The debate on these two accounts has been going
 106 on for several decades.

107 Associative learning mechanisms make sense because
 108 words are typically uttered at the moment when the child
 109 looks at the things that those words refer to. In western
 110 cultures, parents provide linguistic labels of objects for
 111

1 their child when the objects are in the child's visual field.
 Hence, no one doubts that humans can learn cooccurrence
 3 relationships and that the easiest way to teach language is
 to provide linguistic labels of objects at the same moment
 5 that children attend to them. However, parents do not
 carefully name objects for their children in many cultures.
 7 Even in western cultures, words are not always used at the
 moment that their referents are perceived. For instance,
 9 Gleitman [16] showed that most of the time, the child does
 not observe something being opened when the verb "open"
 11 is used. Nevertheless, children have no difficulty in learning
 those words. Associative learning, without further con-
 13 straints or additional information, cannot explain this
 observation.

15 In contrast, the theory of mind reading is able to explain
 many phenomena from the perspective of the inference of a
 17 speaker's referential intentions, especially for the cases
 where words and the corresponding meanings are not
 19 cooccurring, or words are temporally correlated with
 irrelevant meanings. However, the learning environment
 21 in which infants develop does contain statistical regularities
 across words and referents. Meanwhile, empirical studies
 23 (e.g. [43,31]) showed that infants can acquire linguistic
 knowledge based on the statistical properties of speech
 25 input. Taken together, it is very plausible that infants are
 able to acquire the meanings of words based on statistical
 27 regularities from cooccurring words and extralinguistic
 contexts.

29 Fortunately, statistical and social-pragmatic accounts
 are not mutually exclusive. Recently, Hirsh-Pasek et al. [20]
 31 proposed a coalition view in which multiple sources, such
 as perceptual salience, prosodic cue, eye gaze, social
 33 context, syntactic cues, and temporal contiguity, are used
 together by children to learn new words. They argued that
 35 during development, the weighting of individual cues
 changes over time while younger children can just detect
 37 and make use of only a subset of cues in the coalition and
 the older can use a wider subset of cues.

39 The purpose of this study is to show quantitatively the
 effects of both statistical regularities in cross-situational
 41 observation and social cues in multimodal mother-child
 interaction through computational modeling. In early word
 43 learning, children need to start by pairing spoken words
 with the cooccurring possible referents, collecting multiple
 45 such pairs, and then figuring out the common elements.
 Although no one doubts this process, little research has
 47 addressed the details of cross-situational observation. This
 work first introduces a formal model of statistical word
 49 learning which provides a probabilistic framework for
 encoding multiple sources of information. Given multiple
 51 scenes paired with spoken words collected from natural
 interaction between caregivers and their children, the
 53 model is able to compute the association probabilities of
 all the possible word-meaning pairs. Moreover, we argue
 55 that social cues can be naturally integrated in the model as
 additional constraints in statistical computations. The
 57 claim here is that language learners can use social cues,

such as gaze direction, head direction, body movement,
 gesture, intonation of speech, and facial expression, to infer
 59 speakers' referential intentions. We show how these social
 cues can be embodied in the mother's animated actions and
 61 then seamlessly integrated in the framework of statistical
 learning to facilitate word learning. Specifically, we focus
 63 on two kinds of social cues: body movement cues
 indicating the speaker's visual attention and prosodic cues
 65 in speech. This study proposes that those social cues can
 play a spotlight role (shown in Fig. 1) in statistical learning
 67 by leading language learners to focus on the certain aspects
 of a scene and the certain words in speech. Since every
 69 scene is ambiguous and contains multiple possible refer-
 ents, this spotlight function is crucial in solving the word-
 71 to-world mapping problem by making the associative
 process more efficient. The following subsections discuss
 73 how those cues may help in detail.

3.1. The role of deictic body movement in word learning

Ballard et al. [3] argued that at time scales of
 79 approximately one-third of a second, orienting movements
 of the body play a crucial role in cognition and form a
 81 useful computational level, termed the embodiment level.
 At this level, the constraints of the body determine the
 83 nature of cognitive operations. This computation provides
 a language that links external sensory data with internal
 85 cognitive programs and motor actions through a system of
 implicit reference termed deictic, whereby pointing move-
 87 ments of the body are used to bind objects in the world to
 cognitive programs. Examples of sensorimotor primitives
 89 at the embodiment level include an eye movement, a hand
 movement, or a spoken word.

We apply the theory of embodied cognition in the
 context of language learning. To do so, one needs to
 93 consider the role of embodiment from both the perspective

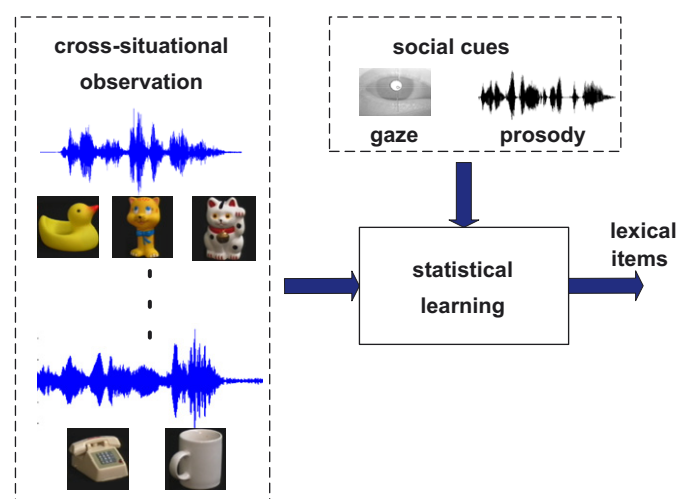


Fig. 1. The basic principle of statistical word learning is to compute
 101 distributional statistics from cross-situational observation. Social cues,
 103 such as joint attention and prosody in speech, can be encoded as
 105 additional constraints in the statistical learning mechanism.
 107
 109
 111
 113

of a speaker (language teacher) and that of a language learner. First of all, recent studies (e.g. [47,28,19]; for review, see [18]) have shown that speech and eye movement are closely linked. Griffin and Bock [19] demonstrated that speakers have a strong tendency to look toward intended referents and words beginning roughly a second after speakers gaze at their referents. Meyer [28] also found that the speakers' eye movements are tightly linked to their speech output. They found that when speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gaze remained on the object until they were about to say the last word about it. Additionally, from the perspective of a language learner, Baldwin et al. [1] showed that infants actively gathered social information to guide their inferences about word meanings and they systematically checked the speaker's gaze to clarify his/her reference. Baldwin et al. [2] proposed that 13-month-old infants give special weight to the cues of indexing the speaker's gaze when determining the referent of a novel label. Their experiments showed that infants established a stable link between the novel label and the target toy only when that label was uttered by an adult who concurrently directed their attention (as indexed by gaze) toward the target. Such a stable mapping was not established when the label was uttered by a speaker who showed no signs of attention to the target toy, even if the object appeared at the same time that the label was uttered and the speaker was touching the object. More recently, Smith [45] suggests that these results may be understood in terms of the child's learning of correlations among actions, gestures, and words of the mature speaker, and intended referents. She argues that construing the problem in this way does not so much "explain away" notions of "mind reading" but rather grounds those notions to the perceptual cues available in the real-time task that infants must solve.

3.2. The role of prosodic cues

When talking to human infants, parents use vocal patterns that are different from normal conversation. They speak slowly and with higher pitch and exaggerated intonation contours. Kuhl [22] argued that the exaggerated prosody was used to define the space of phonemes more clearly for the infant in the 6 to 10 month period where they acquired phonemes. Fernald [15] proposed four developmental functions of intonation in speech to infants. The first function is that infants are attentive to intrinsic perceptual and affective salience in the melodic intonation of mothers' speech. At the second level, the exaggerated intonation patterns of mothers' speech would influence both attentional preference and affective responsiveness of infants. The third function is to allow inferences about speakers' intended meaning from the intonational contours of the utterance. Infants are able to interpret the emotional states of others and make predictions about the future actions of others using information available in vocal and

facial expressions, which provide reliable cues to the affective state and intentions of speakers. The fourth level focuses on the role of prosodic cues in early language development. Fernald argued that the prosody of speech helps to identify linguistic units within the continuous speech signal. Thus, it serves as an attention-focusing device so that mothers use a distinctive prosodic strategy to highlight focused words. Most often, exaggerated pitch peaks are correlated with lexical stress. In light of this, we investigate the role of prosodic cues in early word learning in this paper. Specifically, we focus on the spotlight function of prosody and provide a formal account of how prosodic cues may be used in word learning.

Summarizing all these ideas on embodied cognition, speech production, and social development, the speakers' body movements, such as eye movements, head movements, and hand movements, reveal their referential intentions in verbal utterances, which, in turn almost certainly could play a significant role in early language development [51]. A plausible starting point of learning the meanings of words is the deployment of speakers' intentional body movements to infer their referential intentions. To support this idea, we provide a formal account of how the intentions derived from body movements, which we term *embodied intention*, facilitate statistical word learning. We suggest that infants learn words through their sensitivity to others' intentional body movements in a very specific way: they use spatio-temporal synchrony between speech and referential body movements to find the referents of spoken words.

4. Data

Our study used the video clips of mother–infant interaction from the CHILDES database.¹ These clips contained simultaneous audio and video data wherein a mother introduced her child to a succession of toys stored in a nearby box. The data used for this simulation study were our descriptions of mother–infant interaction. Our description of the audio input—what we fed into the simulated learner—was the entire list of spoken words. Our description of the video stream, again what we fed into the simulated learner, was the list of all objects in view when a word was uttered. Thus, this work assumed that young language learners can (1) segment continuous speech into isolated words (see [21] for a review) and (2) perceive and represent the visual field as basic-level objects (see [13]). In this way, we focused on the mapping problem in word learning—how to associate words to their referents. Specifically, the data was represented as two streams shown in Table 1. The language stream included the transcripts of the mother's speech while the meaning stream consisted of a set of objects as referents. The mother's speech was segmented in spoken utterances based

¹The two video clips we used were contributed by Pamela Rollins. One was labeled as subject di06 and the other one was subject me03.

1 on speech silence. Usually a spoken utterance consisted of
 2 multiple words. Each spoken utterance and the corre-
 3 sponding extralinguistic context formed one learning
 4 situation, while the whole data included multiple such
 5 learning situations. Table 2 shows the statistics of the
 6 training data.

7 In this kind of natural interaction, the vocabulary is rich
 8 and varied and the central items (toy names) are far from
 9 the most frequent words. As shown in Fig. 2, this complex
 10 but perfectly natural situation can be easily quantified by
 11 plotting a histogram of word frequency which shows that
 12 none of the key words—toy names make it into the top 15
 13 items of the list. This entire list of words (those spoken by

the mothers) defines the lexical domain to be learned. An
 elementary idea for improving the ranking of key words
 assumes that the infants are able to weigh the toy
 utterances more by taking advantage of the approximately
 coincident body cues. For instance, the utterances that
 were generated when the infant's gaze was fixated on the
 toys by following the mother's gaze have more weights
 than the ones produced at the moments that the young
 child just looked around while not paying attention to what
 the mother said. We examined the transcripts and weighed
 the words according to how much they were emphasized by
 such cues, but this strategy does little to help spot the toy
 names because those utterances consist of not only toy
 names but also function words, such as *you*, *the*, and *that*,
 which are still the frequent items in the weighted word
 histogram.

We also examined the cooccurrence statistics of word-
 referent pairs as demonstrated in Table 3 (this assumes
 that the learner has completely solved word segmentation
 and morphological analysis problems). Note that only a
 very small set of cooccurring pairs (2.4% and 2.7%) are
 relevant and most of them are irrelevant. For instance, in
 the first line of Table 1, all the words (*oh*, *there*, *went*, *the*,
cow, etc.) could potentially be associated with the
 cooccurring referents (“cow”, “pig”, etc.). Thus, there are
 $6 \times 2 = 12$ possible pairs in this simple learning situation
 and only one pair (*cow* to “cow”) is correct. This example

Table 1
 Examples of spoken utterances and cooccurring extralinguistic contexts

Learning situation	Transcript	Meaning
#1	Oh there went the cow	Cow, pig
#2	There went the cow	Cow, pig
#3	Oh there went the pig	Cow, pig
#4	There we go	Cow, pig
#5	You want to hold the pig	Pig
#6	What is the pig say	Pig
#7	Oinko oinko	Pig
#8	You got pig	Pig
#9	Too much in your hand	Pig

Table 2
 Statistics of data

	Age	Vocabulary	# of total words	# of unique referents	# of total referents	# of utterances
di06	20 mo	336	1236	18	637	281
me03	20 mo	225	1072	21	898	321

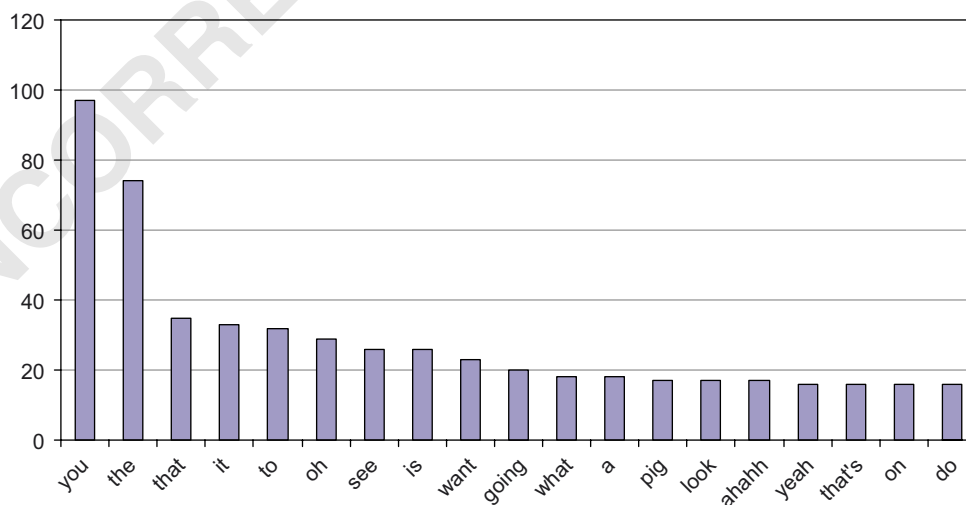


Fig. 2. Word frequency histogram. The histogram of word frequency for subject me03 from Rollins's video data in the CHILDES database shows that the most frequent words are not the central topic meanings.

Table 3
Statistics of cooccurrence of words and referents

	Utterance	Average words per situation	Average referents per situation	Cooccurring pairs	Relevant pairs	Percentage of relevant pairs (%)
di06	281	>4	>2	1414	34	2.4
me03	321	>3	>2	959	26	2.7

Table 4
Most frequently cooccurring word–referent pairs in two video clips

Word	Referent	Frequency	Word	Referent	Frequency
You	Rings	34	Big + bird	Bird	22
You	Rattle	20	A	Bird	18
It	Rings	20	Yeah	Bird	18
The	Rattle	18	The	Bird	17
You	Baby	13	A	Hand	16
To	Rattle	13	You	Bird	16
The	Bird	13	And	Bird	16
The	Pig	12	To	Bird	16
The	Rings	12	Yeah	Hand	15
Pig	Pig	11	There's	Bird	12
That	Baby	10	The	Book	11
On	Rings	10	Look	Bird	10

illustrates the complex learning environment that in which the young children are situated. They need to discover correct word–referent pairs from the environment in which most cooccurring events are irrelevant.

How do they accomplish this task? A further analysis showed that simply selecting word–referent pairs based on cooccurrence frequency cannot achieve the goal. Table 4 lists the most frequent pairs in two video clips. In the left columns for subject me03, only the pair *pig*–“pig”, among 12 frequent pairs, is correct. Similarly, the pair *big + bird*–“bird” is the only relevant pair in the right columns for subject di06. It is observed that cooccurrence statistics tend to associate the most frequent words with the most frequent referents. In natural speech, many function words appear much more frequently than toy names.² In natural interaction, the mother is likely to spend more time on a subset of toys compared with others. Putting these facts together, there are many frequent but irrelevant pairs consisting of function words (in the language stream) and a small set of toys (in the meaning stream). Thus, if young language learners just compute cooccurrence frequencies to build word-to-referent mappings, they would make many wrong associations. Considering the smoothness and efficiency in word learning, it is more likely that infants

²There is evidence that very young children do not attend much to close-class words. Nonetheless, the assumption in this simple analysis is that the learner (at least at the very beginning of word learning) does not have any linguistic knowledge. Starting with this assumption, the present study shows how a statistical mechanism can induce correct word–referent pairs based on cross-situational observations. The model also discovers that the high cooccurrence of function words with different referents makes them irrelevant to almost any single referent.

utilize a more effective strategy to deal with the reference uncertainty problem.

5. A statistical model of cross-situational observation

Our solution to the computational problem of word learning rests on advances in machine translation. Briefly, machines “learn” word correspondences (which word in one language corresponds to which word in another language) by finding the statistical regularities across large parallel corpora in two languages. Here, we use this same computational approach but conceptualize the video stream as one language and the audio stream as the other. Associating meanings (toys, etc.) with words (toy names, etc.) can be viewed as the problem of identifying word correspondences between English and the meaning language. This conceptualization provides a unique way to understand statistical learning of word-to-world mapping. With this perspective, we develop our model based on the translation model proposed in [8]. The central idea is that word–meaning pairs are latent variables underneath the observations that consist of spoken words and extralinguistic contexts. Thus, association probabilities of these pairs are not directly observable, but they somehow determine the observations because spoken language is produced based on the mother’s lexical knowledge. Therefore, the objective of young language learners or computational models is to figure out the values of these underlying association probabilities so that they can increase the chance of obtaining the observations. Correct word–meaning pairs are those which can maximize the likelihood of the audio-visual observations in natural interaction.

In practice, the learning process can be formalized as an expectation–maximization algorithm (EM) [12]. The data of two languages in parallel (English and extralinguistic meaning language) can be treated as a probability distribution. The idea of EM is to represent the data as the sum of component probability distributions. More specifically, the probability of each word is expressed as a weighted mixture consisting of the conditional probabilities of each word given its possible meanings. The task of the learning algorithm is then to find the reliable associations of object names and their meanings which maximize the likelihood function of observing the whole data set.

The general setting is as follows: suppose we have a word set $X = \{w_1, w_2, \dots, w_N\}$ and a meaning set $Y = \{m_1, m_2, \dots, m_M\}$, where N is the number of words and M is the number of meanings (toys, etc.). Let S be the

1 number of learning situations. All word data are in a set
 $\chi = \{(S_w^{(s)}, S_m^{(s)}), 1 \leq s \leq S\}$, where for each learning situation,
 $S_w^{(s)}$ consists of r words $w_{u(1)}, w_{u(2)}, \dots, w_{u(r)}$, and $u(i)$ can be
 selected from 1 to N . Similarly, the corresponding
 contextual information $S_m^{(s)}$ in that learning situation
 include l possible meanings $m_{v(1)}, m_{v(2)}, \dots, m_{v(l)}$ and the
 value of $v(j)$ is from 1 to M . A simple example in Fig. 3
 consists of two learning situations in which every word can
 potentially be associated with any cooccurring meaning.
 The computational challenge here is to build several one-
 to-one mappings (e.g. *cow* to “cow”) from many-to-many
 possible associations. We suggest that to figure out which
 word goes to which meaning, language learners do not
 consider the association of just a single word–referent pair,
 but they estimate all these possible associations simulta-
 neously. Thus, they attempt to estimate the association
 probabilities of all of these pairs so that the best overall
 mapping can be achieved. In doing so, the constraints
 across multiple learning situations and the constraints
 across different word–referent pairs are jointly considered
 in a general system which attempts to discover the best
 translation between words and referents based on statisti-
 cal regularities in the observation.

Formally, given a data set χ , we use the machine
 translation method proposed by Brown et al. [8] to
 maximize the likelihood of generating the meaning strings
 given English descriptions:

$$\begin{aligned}
 P(S_m^{(1)}, S_m^{(2)}, \dots, S_m^{(S)} | S_w^{(1)}, S_w^{(2)}, \dots, S_w^{(S)}) \\
 &= \prod_{s=1}^S \sum_a p(S_m^{(s)}, a | S_w^{(s)}) \\
 &= \prod_{s=1}^S \frac{\varepsilon}{(r+1)^l} \prod_{j=1}^l \sum_{i=0}^r p(m_{v(j)} | w_{u(i)}), \quad (1)
 \end{aligned}$$

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

To maximize the above likelihood function, a new
 variable $c(m_m | w_n, S_m^{(s)}, S_w^{(s)})$ is introduced which represents
 the expected number of times that any particular word w_n
 in a language string $S_w^{(s)}$ generates any specific meaning m_m
 in the cooccurring meaning string $S_m^{(s)}$:

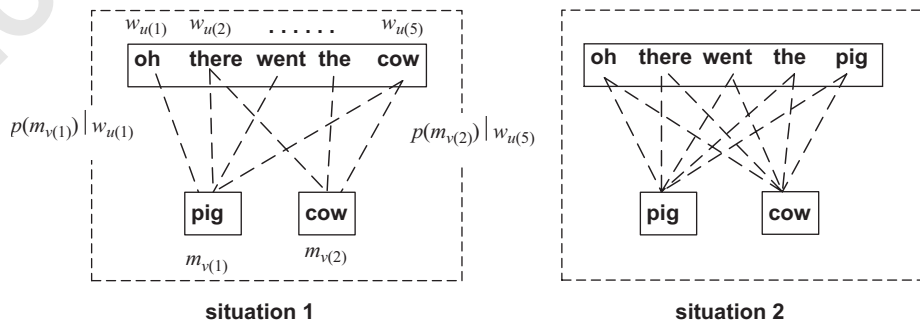


Fig. 3. Every word–referent pair could be potentially relevant. The task of the model is to estimate the possible links between words and referents.

$$\begin{aligned}
 c(m_m | w_n, S_m^{(s)}, S_w^{(s)}) &= \frac{p(m_m | w_n)}{p(m_m | w_{u(1)}) + \dots + p(m_m | w_{u(r)})} \\
 &\times \sum_{j=1}^l \delta(m_m, v(j)) \sum_{i=1}^r \delta(w_n, u(i)), \quad (2)
 \end{aligned}$$

where δ is equal to one when both of its arguments are the
 same and equal to zero otherwise. The second part in Eq.
 (2) counts the number of cooccurring times of w_n and m_m .
 The first part assigns a weight to this count by considering
 it in the context of all the other words in the same learning
 situation. By introducing this new variable, the computa-
 tion of the derivative of the likelihood function (shown in
 Eq. (1)) with respect to the association probability
 $p(m_m | w_n)$ results in:

$$p(m_m | w_n) = \frac{\sum_{s=1}^S c(m_m | w_n, S_m^{(s)}, S_w^{(s)})}{\sum_{m=1}^M \sum_{s=1}^S c(m_m | w_n, S_m^{(s)}, S_w^{(s)})}. \quad (3)$$

Algorithm 1. Estimating word–referent association prob-
 abilities

Assign initial values for $p(m_m | w_n)$ based on cooccurrence
 statistics.

repeat

E-step: Compute the counts for all word–referent
 pairs using Eq. (2).

M-step: Reestimate the association probabilities using
 Eq. (3).

until the association probabilities converge.

As shown in Algorithm 1, the method sets an initial
 $p(m_m | w_n)$ to be flat distribution, and then successively
 computes the occurrences of all word–meaning pairs
 $c(m_m | w_n, S_m^{(s)}, S_w^{(s)})$ using Eq. (2) and the association
 probabilities using Eq. (3). In this way, our method runs
 multiple times and allows for reestimating word–referent
 association probabilities. In practice, in addition to objects
 in the scene, a special referent “NON” is also added in each
 meaning stream. The hypothesis is that language learners
 are aware that some words may not have referents in the
 extralinguistic context and that consequently those words
 should be associated with a meaning that does not refer to
 anything in the visual field. That said, the model is also able
 to discover these words (function words, etc.) that do not

1 have the concrete semantic meanings based on their
2 association probabilities to “NON”. The detailed technical
3 descriptions can be found in [8,52].

4 The central idea of the algorithm can be illustrated with
5 a simple toy example shown in Table 5. The input to the
6 simulated learner is rather simple, consisting of four
7 spoken utterances (eight words in total) and two objects.
8 This kind of input makes statistical learning hard because
9 there are fewer regularities that the model can utilize.
10 Nonetheless, the example serves to demonstrate how the
11 learning algorithm performs and what statistical regula-
12 rities can be obtained from such limited data. Table 6
13 shows the results wherein the third column is the
14 cooccurrence statistics of word–referent pairs from which
15 we cannot separate correct pairs from irrelevant ones. The
16 fourth and fifth columns show association probabilities of
17 all the cooccurring pairs. Clearly, with the convergence of
18 the algorithm, the association probabilities of most correct
19 pairs increase, such as *dog* to “dog”, *cat* to “cat”, and *here*
20 to “NON”. Accordingly, the association probabilities of
21 irrelevant pairs decrease. Also note that for the words that
22 just appear once in the data, the EM algorithm has

23 Table 5
24 A simple example

Transcript	Referent
Here is a cat	Cat, dog, NON
Here is a dog	Dog, NON
You like cat	Cat, NON
Here is it	NON

33 Table 6
34 A simple example

Word	Referent	# of Cooccurrence	Association probability iteration 1	Association probability iteration 2
Here (3)	NON (4)	3	0.4821	0.5387
Here	Cat (3)	1	0.1793	0.1282
Here	Dog (2)	2	0.3386	0.3331
Cat (2)	NON	2	0.4	0.3747
Cat	Cat	2	0.4	0.5105
Cat	Dog	1	0.2	0.1148
Dog (1)	NON	1	0.5	0.4582
Dog	Dog	1	0.5	0.5418
A (2)	NON	2	0.3953	0.3597
A	Cat	1	0.2093	0.1779
A	Dog	2	0.3953	0.4624
Is (2)	NON	3	0.4821	0.5387
Is	Cat	1	0.1793	0.1282
Is	Dog	2	0.3386	0.3331
You (1)	NON	1	0.5	0.5
You	Dog	1	0.5	0.5
Like (1)	NON	1	0.5	0.5
Like	Dog	1	0.5	0.5
It (1)	NON	1	1	1

55 The numbers in parenthesis indicate the number of occurrence of each
56 item (word or referent).

25 relatively little impact on reestimating their association
26 probabilities. With such a small amount of data, the
27 example demonstrates that the EM algorithm can poten-
28 tially discover correct word–referent pairs in the data that
29 are not visible to a naive eye. This ability lies in the fact
30 that it computes association probabilities of all the pairs by
31 considering them as a system of words that interact with
32 and influence each other, but not by just counting
33 cooccurrences of individual pairs.

34 Two measures were used to evaluate the performance on
35 the infant data: (1) word–meaning association accuracy
36 (precision) measures the percentage of the words spotted
37 by the model which are actually correct and (2) lexical
38 spotting accuracy (recall) measures the percentage of
39 correct words that the model learned among all the
40 relevant words that are expected to be learnt. Fig. 4 shows
41 that the model of statistical associative learning strikingly
42 improves the association probabilities of the word–referent
43 pairs compared with the method of counting cooccurrence
44 statistics. Seventy-five per cent of words are associated with
45 correct meanings, such as the word *hat* paired with the
46 meaning “hat” (in purple color in the figure) and the word
47 *book* paired with the meaning “book” (in green color in the
48 figure). In addition, all the toy words are in the top three
49 most relevant words of the corresponding objects (col-
50 umns). The overall recall accuracy is 58%. Note that the
51 object “ring” seems to relate to multiple words. That is
52 because in the video clips, the mothers introduced to the
53 children to a set of rings with different colors. Therefore,
54 they spent significantly more time on the object “ring” and
55 consequently many words cooccur more frequently with
56 the meaning “ring” compared with other meanings. This
57 result may seem counterintuitive according to a cross-
58 situational perspective because generally the more time the
59 mother spends on the toy, the more likely the young child
60 can learn the name of this toy. However, a detailed
61 statistical analysis, such as the model described in this
62 paper, indicates that distributional statistics across words,
63 across referents, and across the cooccurrences of the items
64 in these two streams jointly determine whether a word–re-
65 ferent pair is easy to discover. The ability to disambiguate
66 many-to-many cooccurrences to build one-to-one map-
67 pings lies in not only the statistics of a single word, but also
68 how this word is distributed across different contexts and
69 what words surround it in different contexts. Considering
70 that in an extreme case, if there is only one object in the
71 scene and the mother produces many words when she
72 introduces the toy to the child, then statistically there is not
73 enough information to determine among all these words
74 which one is associated with the toy. More generally,
75 statistical language learning needs a sufficient amount of
76 cross-situational observation including multiple contexts
77 and multiple words in these contexts to discover the
78 statistically reliable associations. In this experiment, since
79 the simulated learner is input with limited data, several
80 words (including both the word “ring” and other irrelevant
81 words) cooccur relatively frequently with the referent

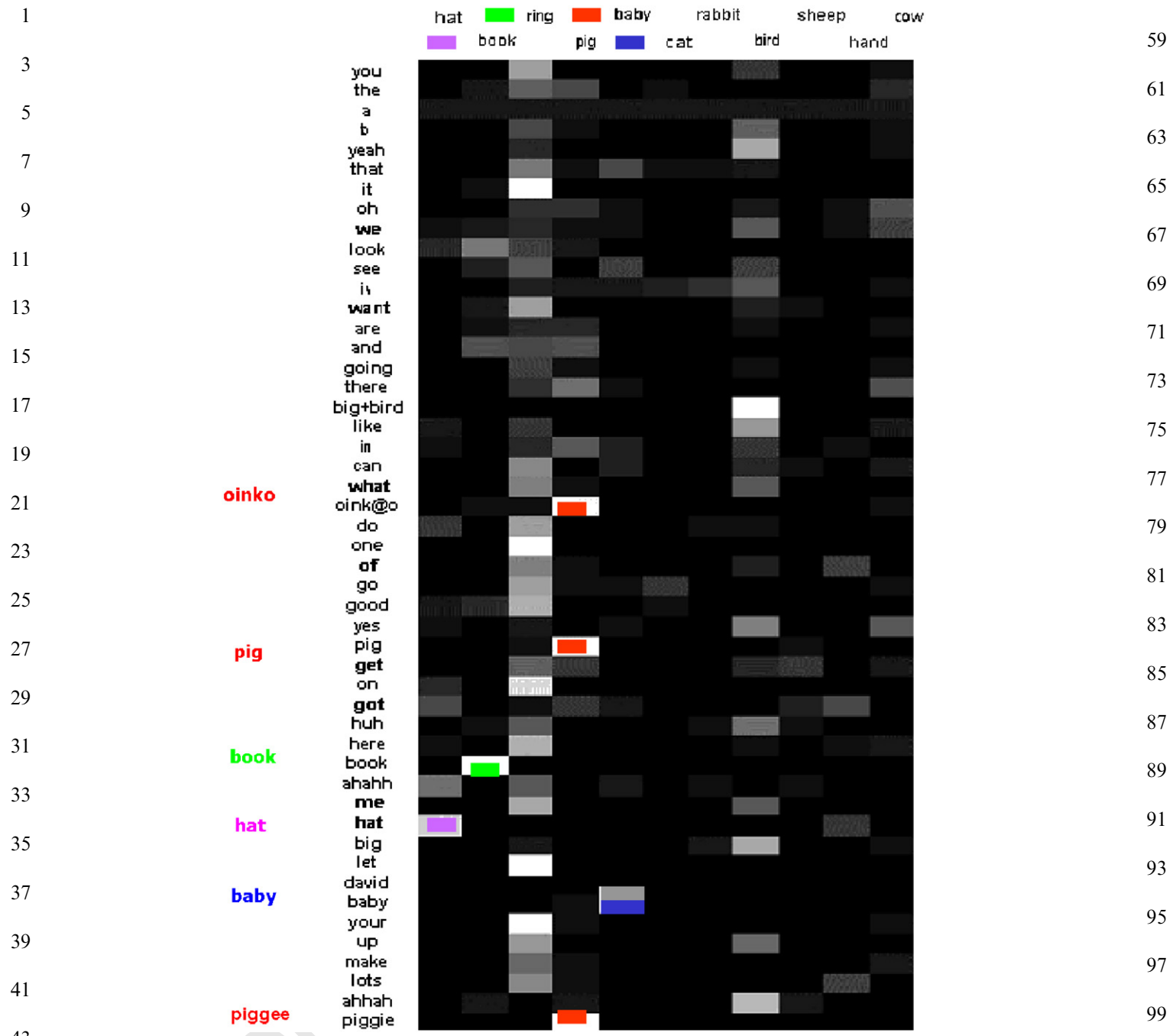


Fig. 4. The results of statistical word learning to build word-to-world mappings. The row is a list of words and the column is a list of meanings. Each cell is the association probability of a specific word-meaning pair. Dark color means low probability while white means high probability.

“ring” and rarely occur in any other context. Therefore, as a purely statistical observer and without any other information (such as social cues discussed in the next experiment), the model determines to assign high association probabilities between those words and the “ring” because this is the statistically best solution based on limited data.

It is also interesting that *eye* is associated with the meaning “bird”. Strictly speaking, the word *eye* is not relevant to the object “bird”. However, our simulated

learner observes that these two things not only cooccur in time but also are likely to be associated based on distributional statistics in the data. Our explanation of this result is that the performance of statistical learning is based on the amount of the data fed in the model. The more data are used, the more statistical regularities are in the learning environment, and the more likely the model is able to discover those regularities and work in a more efficient way. Furthermore, statistical word learning is a cumulative procedure. Young language learners acquire

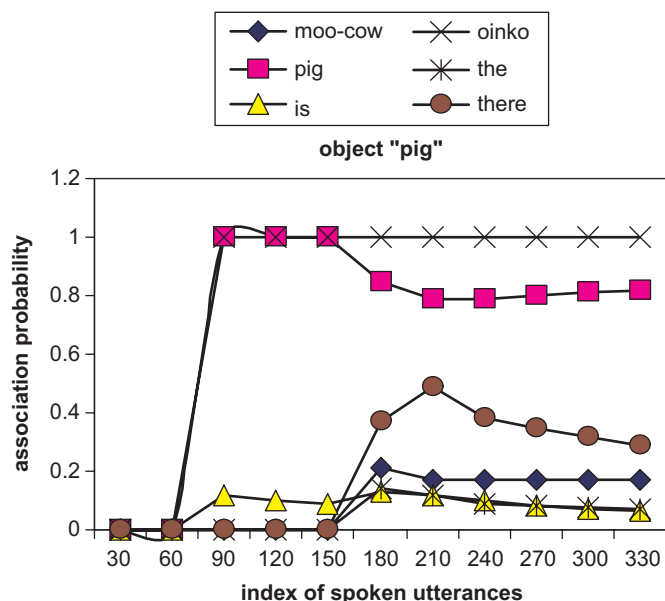


Fig. 5. Word–meaning pairs in the course of refinement by our EM algorithm. The word *pig* and the word “oinko” are continuously likely to associate with the meaning “pig”.

new words day by day. During this developmental process, they might build several hypotheses, such as both *eye* and *bird* associated with the meaning “bird”, because both words are highly correlated in time with the same meaning “bird”. With our statistical cross-situational learning view, when they hear the word *eye* in other contexts, for example, with other animals or humans (but not “bird”), they will learn that the word *eye* and the word *bird* have different meanings and that *eye* should be associated with the meaning “eye”. With this perspective, we treat *eye* as a good instance based on the data utilized so far. These kinds of associations, in our view, are partial knowledge and serve as precursors for correct associations during development, which will lead to a perfect association when the word is perceived again in different contexts. Fig. 5 shows the results of word learning from an incremental view and provides a detailed description of the dynamics of our learning model. The model starts with limited input and the estimates of association probability are not reliable. Specifically, some irrelevant word–meaning pairs obtain relatively high probabilities. With more training data, the model converges to more accurate estimates.

So far we have shown that a statistical model can compute the association probabilities of all the cooccurring word–meaning pairs in the data. Moreover, this formal model provides a probabilistic framework for studying the role of other factors and constraints in word learning, such as social cues and syntactic constraints.

6. The integration of social cues in statistical learning

The communication between infants and their caregivers is multisensory, which involves seeing, hearing, touching,

and pointing. This paper argues that social cues encoded in multimodal interaction highlight target word–referent relations for young language learners. In a bidirectional relationship between maternal multimodal communication styles and infants’ perception of word–referent relations, mothers synchronize their verbal references and non-verbal body movements (eye gaze, gesture, etc.) to highlight word–referent relations for young children. That is, presenting information across multiple modalities simultaneously serves to highlight the relations between the two patterns of stimulation. Also, from the learner’s perspective, infants are able to rely on observing mothers’ eye gaze and other pointing motions to detect their intended referents. Thus, both mothers and infants actively use multimodal communication to solve the mapping problem in lexical acquisition. This study provides a quantitative account of how those multimodal cues can facilitate word learning. Specifically, we focus on two social cues: joint-attention cues as deictic reference and prosodic cues in maternal speech. We suggest that these social cues serve as spotlights to highlight a subset of words in the speech stream and a subset of objects in the meaning stream, respectively (shown in Fig. 6).

6.1. Visual spotlight

Children as young as 12–18 months spontaneously check where a speaker is looking when he/she utters a word, and then link the word with the object the speaker is looking at. This observation indicates that joint visual attention (deictic gaze) is a critical factor in development and learning [11]. In a natural (and most often complex) learning environment, this visual spotlight gives maximal processing to the specific part of the visual field. During natural infant–caregiver interaction, joint visual attention involves detecting body cues that indicate the direction of a caregiver’s attention to the object in the scene, and then moving the body, head, and eyes to acquire the target object with high-resolution focal vision, which is one of the crucial steps to deal with the word-to-world mapping problem.

As shown in Table 7, we used two categories to describe extralinguistic contextual information for each learning situation (defined by a spoken utterance based on speech silence). One category consists of the objects of joint attention by the child and the mother. The second represents all the other objects in the visual field. Fig. 6 illustrates two examples of speech–scene pairs in which the shaded meanings are attended objects and non-shaded meanings are other objects in the scene. In Section 6.3, we describe the method that makes use of this joint-attention information in word learning.

6.2. Prosodic spotlight

Snedeker and Trueswell [46] showed that speakers produce and listeners use prosodic cues to disambiguate

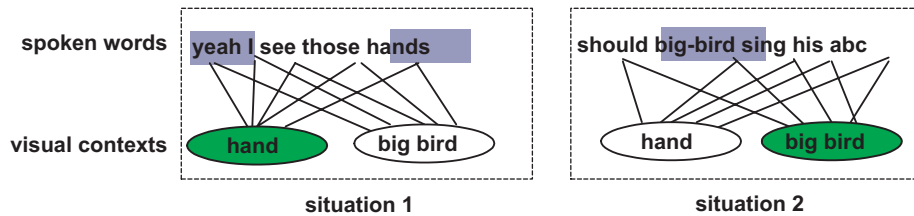


Fig. 6. Cross-situational word-meaning association with social cues. The prosodic cues highlight some words in speech and the cues of joint attention highlight attended objects in visual contexts.

Table 7
Examples of transcriptions and contextual labels

Transcriptions	Attended objects	Other objects
The kitty-cat go meow meow	Kitty-cat	Baby, big-bird, rattle, book
Ah and a baby	Baby	Kitty-cat, big-bird, rattle, book
There's a baby just like my David	Baby	Kitty-cat, big-bird, rattle, book
A baby	Baby	Kitty-cat, big-bird, rattle, book
That's a nice book	Book	Kitty-cat, big-bird

alternative meanings of a syntactic phrase in a referential communication task, suggesting that in addition to linguistic information (what is said, etc.), non-linguistic aspects of speech (how to say it, etc.) also contains important information. In fact, research on infant-directed speech (e.g. [15]) showed that mothers adapt their verbal communication to infants in order to facilitate their language learning. In light of this, we analyzed maternal speech by extracting low-level acoustic features and using those features to spot the words emphasized by language teachers. We argue that perceptually salient prosodic patterns may serve as spotlights on linguistic information conveyed by speech. One role of prosodic cues in word learning, we suggest, is to help young learners identify key words from the speech stream.

In a natural toy-play session, we argue that prosodically salient words in maternal speech serve two basic functions with respect to word learning and can be categorized into two classes accordingly. The first group of words serve as communication of intention and emotion. An important role of those words is to attract the child's attention so that she would follow what the mother talks about and what she looks at. In this way, both the mother and the language learner share attention, which is a cornerstone in social and language development. The right column in Fig. 7 illustrates an example in the video clips in which the mother used high pitch on the word *you* to attract the child's attention. Some other common words and phrases frequently used by the mother are *yeah*, *oh*, *look*, and *that's*. The other group of words contain the most important linguistic information that the mother intends to convey. In the toy-play sessions used in this study, most of those

words refer to the concepts that are related to visual objects in the physical environment, such as object names, their colors, sizes, and functions. An example of the words in the second group is the object name *baby* shown in the left column of Fig. 7.

In implementation, CMU sphinx speech recognition system was used to align maternal speech and transcriptions [24]. As a result, the timestamps of the beginning and end of each spoken word were extracted. Next, we made three kinds of low-level acoustic measurements on both an utterance and the words embedded in the utterance. The following prosodic features were extracted based on pitch (fundamental frequency f_0):

- *75 percentile pitch* p_{75} : The 75 percentile pitch value of all voiced parts of the speech unit.
- *Delta pitch range* p_r : The change in pitch between frames (20 ms) was calculated as delta pitch. This measure represents the difference between the highest and the lowest delta pitch values within the unit (utterance or word).
- *Mean delta pitch* p_m : The mean delta pitch of the voiced part of the spoken unit.

For each feature, we extracted the values over both an utterance and the words within the utterance to obtain the prosodically highlighted words in each spoken utterance. To do so, we compared the extracted features of a word with those features of the utterance containing that word, which indicates whether a word sounds "highlighted" in the acoustic context. Specifically, for the word w_i in the spoken utterance u_j , we formed a feature vector: $[p_{75}^{w_i} - p_{75}^{u_j}, p_r^{w_i} - p_r^{u_j}, p_m^{w_i} - p_m^{u_j}]^T$, where $p_m^{u_j}$ is the mean delta pitch of the utterance and $p_m^{w_i}$ is that of the word. In this way, the prosodic envelope of a word is represented by a three-dimensional feature vector. We used the support vector clustering (SVC) method [5] to group data points into two categories. One consists of prosodically salient words and the other one includes non-emphasized words. In the SVC algorithm, data points are mapped from the data space to a high-dimensional feature space using a Gaussian kernel. In this feature space, the algorithm looks for the smallest sphere that encloses the data, and then maps the data points back to the data space and forms a set of contours to enclose them. These contours can be interpreted as cluster boundaries.

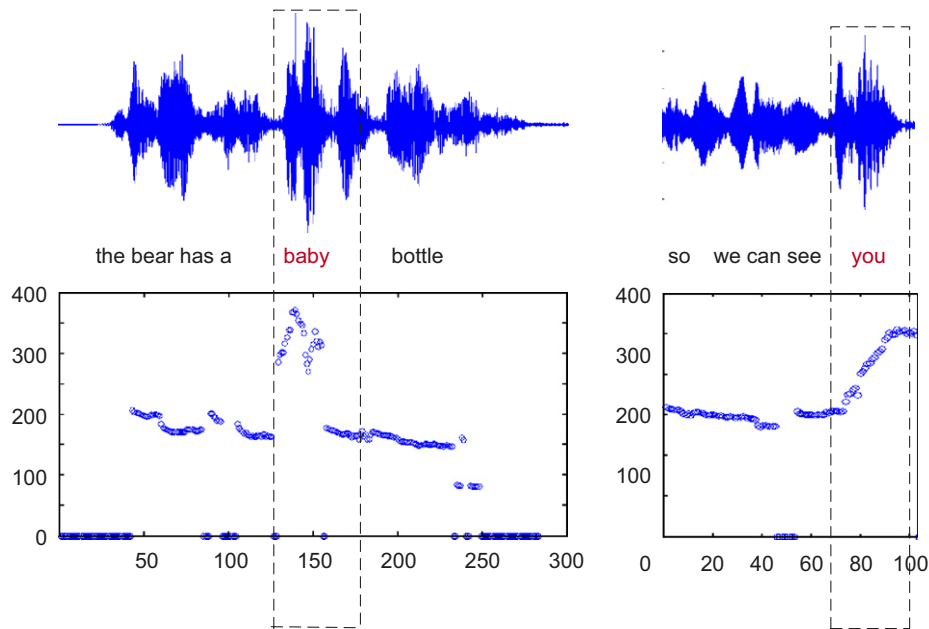


Fig. 7. Speech and intonation. The prosodic cues highlight several words. The first column represents speech signals and the second column shows the profiles of fundamental frequency (f_0). The word *baby* is highlighted in the left utterance and the word *you* is prosodically distinctive from others in the right utterance.

6.3. Modeling the role of social cues in statistical learning

Our approach to encoding social cues in the framework of the statistical learning model is to give weights to both words in the speech stream and referents in the meaning stream based on whether they are spotlighted by social cues. The hypothesis is that the objects in joint attention are more likely referred in speech. Similarly, the prosodically distinctive words are more likely related to the toys in the interaction. Formally, each word $u(i)$ is assigned with a weight $w_p(i)$ based on its prosodic category. Similarly, each visual object $v(j)$ is set with a weight $w_v(j)$ based on whether it is attended by the speaker and the learner. In this way, the same method described in the previous section is applied and the only difference is that the estimate of $c(m_m|w_n, S_m^{(s)}, S_w^{(s)})$ is now given by

$$c(m_m|w_n, S_m^{(s)}, S_w^{(s)}) = \frac{p(m_m|w_n)}{p(m_m|w_{u(1)}) + \dots + p(m_m|w_{u(r)})} \times \sum_{j=1}^l \delta(m_m, v(j)) * w_v(j) \times \sum_{i=1}^r \delta(w_n, u(i)) * w_p(i). \quad (4)$$

In practice, we set the predetermined values of $w_v(j)$ and $w_p(i)$ to be 3 for highlighted objects and words. The weights of all the other words and objects are set to be 1.

6.4. Experimental results

Four methods were applied on the same data and the results of precision and recall (defined in Section 5) are as

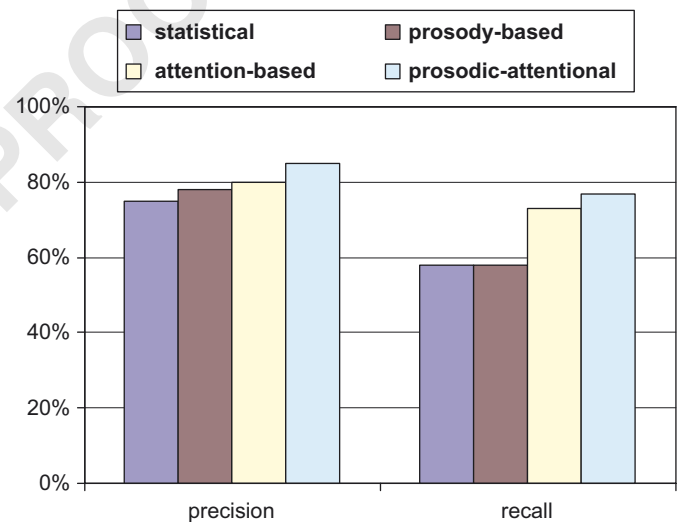


Fig. 8. A comparative study of four methods. Precision and recall are two metrics to measure how well the methods build correct word-referent pairs.

follows (also shown in Fig. 8): (1) purely statistical learning (75% and 58%), (2) statistical learning with prosodic cues (78% and 58%), (3) statistical learning with the cues from visual attention (80% and 73%), and (4) statistical learning with both attentional and prosodic cues (83% and 77%). Fig. 9 shows the comparative results of these four approaches on specific instances. Ideally, association probabilities of the first or second words are expected to be high and others to be low. For instance, the first plot represents the meaning of the object “cat”. Both the spoken word *kitty-cat* and the spoken word *meow* are closely relevant to this meaning. Therefore, the association

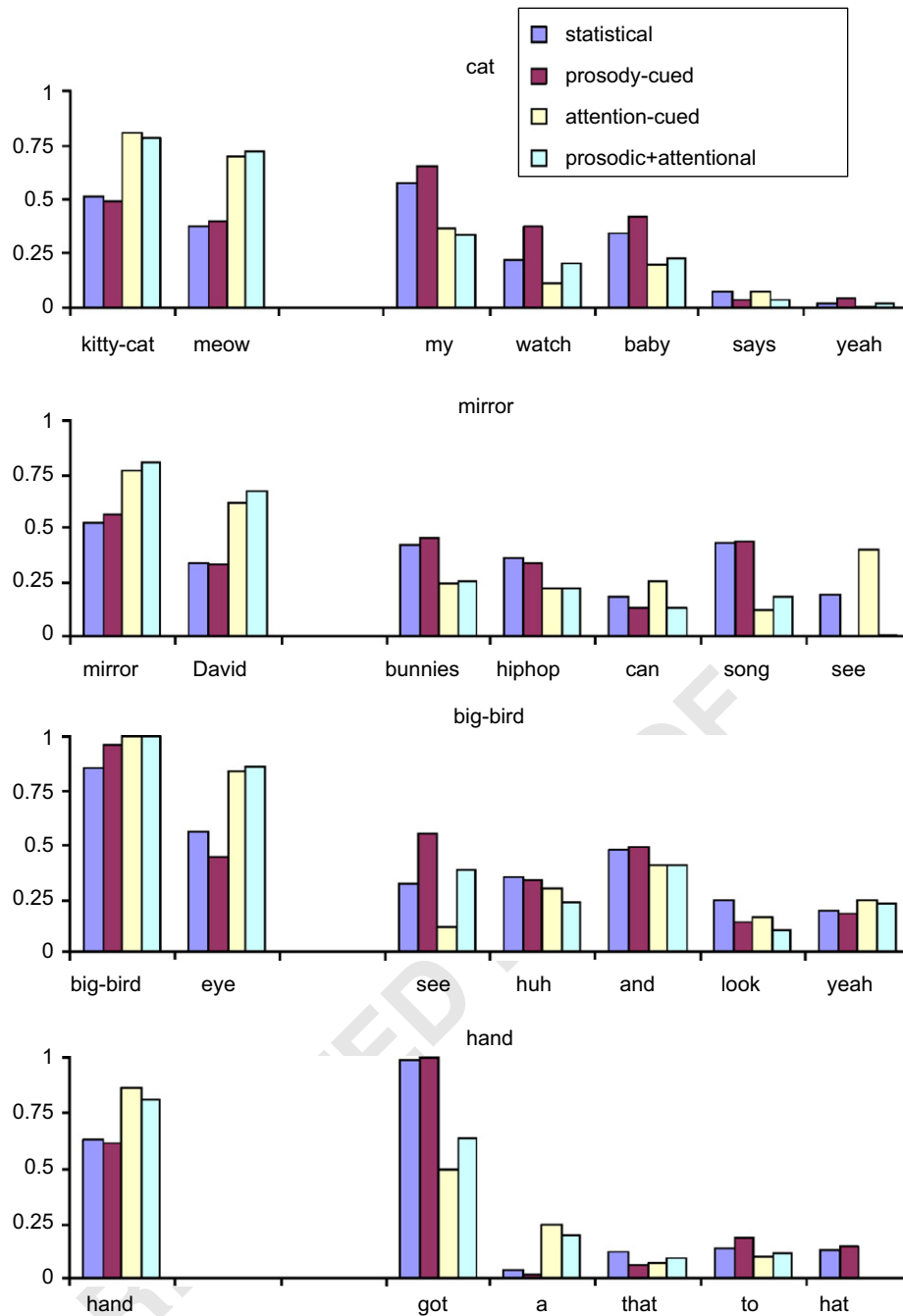


Fig. 9. The comparative results of the methods considering different sets of cues. Each plot shows the association probabilities of several words to one specific meaning labeled on the top. The first one or two items are correct words that are relevant to the meanings and the following words are irrelevant.

probabilities are high for these two words and are low for all the others words, such as *my*, *watch*, and *baby*, which are not correlated with this context. Note that in the meaning of the object “bird”, we count the word *eye* as a positive one because the mother uttered it several times during the interaction when she presented the object “bird” to her child. Similarly, when she introduced the object “mirror”, she also mentioned the child’s name *David* whose face appeared in the mirror.

An approach that just randomly maps words to available referents by chance can obtain only 5.3% precision and 15.2% recall accuracy. Compared with that approach, the results of the statistical learning approach (the first bars) are reasonably good. For instance, it obtains *big-bird* and *eye* for the meaning “bird”, *kitty-cat* for the meaning “cat”, *mirror* for the meaning “mirror”, and *hand* for the meaning “hand”. But it also makes wrong estimates, such as *my* for the meaning “cat” and *got* for the meaning “hand”. We expect that attentional and

prosodic constraints will make the association probabilities of correct words higher and decrease the association probabilities of irrelevant word–meaning pairs. The method encoding prosodic cues moves toward this goal although occasionally it changes the probabilities in the reverse way, such as increasing the probability of *my* to the meaning “cat”. What is really helpful is to encode the cues of joint attention. The attention-cued method significantly improves the accuracy of estimate for almost every word–meaning pair. Of course, the method including both joint-attention and prosodic cues achieves the best performance. Compared with purely statistical learning, this method highlights the correct associations (e.g. *kitty-cat* with the meaning “cat”), and decreases the irrelevant associations, such as *got* with the meaning “hand”. In this method, we can simply select a threshold and pick the word–meaning pairs which are overlapped with the majority of words in the target set. We need to point out that the results here are obtained from very limited data. Without any prior knowledge of the language (the most challenging case in word learning), the model is able to learn a significant number of correct word–meaning associations. It is also worth noting that not all “errors” are irrelevant to the meaning of the target words. For example, the fact that the meaning “hand” is statistically related to the verb *got* is interesting, in that the meaning of “getting” in many contexts—and perhaps to children—may include the taking of an object in hand.

A further analysis shown in Fig. 10 provides a more detailed description of the roles of different cues in word learning, suggesting that the gain from adding social cues is significant. The left bars show the average of association probabilities of correct word–referent pairs and the right bars show the average of association probabilities of the first three irrelevant pairs. The reason that only three irrelevant words are considered for each object is that for a word, its association probabilities to most referents are close to zero. This is due to the fact that in most cases, the word just occurs in the subset of contexts but not all of them. Therefore, the average of the association probabilities of the first three irrelevant pairs is more informative

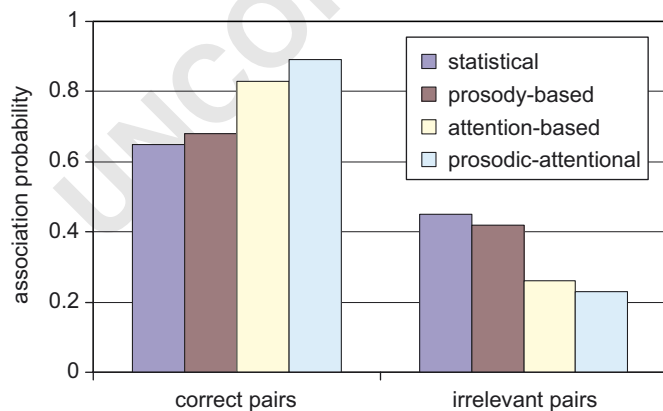


Fig. 10. The improvement of association probabilities.

than that of all the pairs. The figure shows that encoding social cues in statistical learning not only improves the overall results in terms of precision and recall but also makes the association probabilities of correct word–referent pairs to increase and meanwhile association probabilities of irrelevant pairs to decrease. Furthermore, the difference between the average association probabilities of correct pairs and that of irrelevant ones is also a good indicator whether the model can easily set a threshold of association probability to distinguish the correct pairs from others.

7. Conclusion

To learn words in their native language, young learners need to address at least three problems. First, they need to segment continuous speech into isolated words. This problem is difficult because spoken language lacks the acoustic analog of blank spaces in text. Second, they need to extract the possible meanings of these words from extralinguistic contexts. This procedure involves encoding and storing sensorimotor experiences in the brain to build prelinguistic concepts. Third, they need to associate these concepts with linguistic labels. Our previous work [52] presented a computational model that is able to discover spoken words from continuous speech and associate them with their perceptually grounded meanings. That model tackled the three tasks and provided a more complete picture of early word learning. Nonetheless, Gleitman [17] argued that a major problem in word acquisition is not about how to represent the concepts to form the meanings of words but a considerable part of the bottleneck resides in the (either innate or learned) mechanisms, constraints, and tools that can be utilized to solve the mapping problem (the third task). In an effort to make concrete progress, we made some assumptions in this modeling work which allow us to focus on the mapping problem—the most fundamental problem in lexical acquisition. Thus, the simulated learner here already knows how to acquire individual words and meanings and only needs to deal with the word-to-world mapping problem.

We believe that in natural infant–caregiver interaction, the mother provides non-linguistic signals to the infant through her body movements, the direction of her gaze, and the timing of her affective cues via prosody. Previous experiments have shown that some of these non-linguistic signals can play a critical role in infant word learning, but a detailed estimate of their relative weights has not been provided. Based on statistical learning and social-pragmatic theories, this work proposed a unified model of early word learning, which integrates statistical and social cues to enable the word-learning process to function effectively and efficiently. In our model, we explored the computational role of non-linguistic information, such as joint attention and prosody in speech, and provided the quantitative results to compare the effects of different statistical and social cues. We need to point out that the

1 current unified model does not encode any syntactic
 3 properties of the language, which definitely play a
 5 significant role in word learning, especially in the later
 7 stage. Therefore, one natural extension of the current work
 9 is to add the syntactic constraints in the current probabilistic
 11 framework to study how this knowledge can help the
 13 lexical acquisition process and how multiple sources can be
 15 integrated in a general system.

17 Also, the weighing mechanism to encode social cues in
 19 statistical word learning is straightforward and rather
 21 simple, which serves as first step towards understanding the
 23 integration and interaction of statistical and social cues.
 25 One direction of future work is to investigate the
 dependencies of these cues because the mother is able to
 use multimodal cues simultaneously and effectively. For
 instance, our results showed that prosodic cues seem to be
 not very informative in word learning because the mother
 used them to not only highlight the linguistic information
 but also to attract the child's attention. However, if the
 child and the mother were already in joint-attention state,
 then prosodic cues might most often be used to highlight
 the linguistic content. Thus, we will move beyond the study
 of individual social cues and study how they coordinate in
 multimodal interaction.

27 References

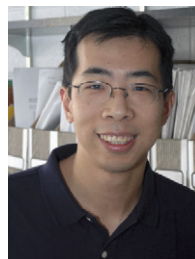
29 [1] D. Baldwin, Early referential understanding: infant's ability to
 31 recognize referential acts for what they are, *Dev. Psychol.* 29 (1993)
 832–843.
 [2] D.A. Baldwin, E.M. Markman, B. Bill, R.N. Desjardins, J.M. Irwin,
 G. Tidball, Infant's reliance on a social criterion for establishing
 word-object relations, *Child Dev.* 67 (1996) 3135–3153.
 [3] D.H. Ballard, M.M. Hayhoe, P.K. Pook, R.P.N. Rao, Deictic codes
 for the embodiment of cognition, *Behav. Brain Sci.* 20 (1997)
 1311–1328 URL: ([citeseer.nj.nec.com/
 ballard95deictic.html](http://citeseer.nj.nec.com/ballard95deictic.html)).
 [4] S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of
 Mind*, The MIT Press, Cambridge, MA, 1995.
 [5] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, Support vector
 clustering, *J. Mach. Learn. Res.* 2 (2001) 125–137.
 [6] P. Bloom, Intentionality and word learning, *Trends Cognit. Sci.* 1 (1)
 (1997) 9–12.
 [7] P. Bloom, *How Children Learn the Meanings of Words*, The MIT
 Press, Cambridge, MA, 2000.
 [8] P.F. Brown, S. Pietra, V. Pietra, R.L. Mercer, The mathematics of
 statistical machine translation: parameter estimation, *Comput.
 Linguist.* 19 (2) (1994) 263–311.
 [9] G. Butterworth, The ontogeny and phylogeny of joint visual
 attention, in: A. Whiten (Ed.), *Natural Theories of Mind: Evolution,
 Development, and Simulation of Everyday Mindreading*, Blackwell,
 Oxford, UK, 1991, pp. 223–232.
 [10] E. Colunga, L. Smith, A connectionist account of the object-sub-
 stance distinction, *Psychol. Rev.* 112 (2005) 347–382.
 [11] G. Deak, J. Triesch, The emergence of attention-sharing skills in
 human infants, in: K. Fujita, S. Itakura (Eds.), *Diversity of
 Cognition*, University of Kyoto Press, in press.
 [12] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from
 incomplete data via the em algorithm, *J. R. Statist. Soc.* 39 (1) (1977)
 1–38.
 [13] P. Eimas, P. Quinn, Studies on the formation of perceptually-based
 basic-level categories in young infants, *Child Dev.* 65 (1994) 903–917.

[14] J. Elman, E. Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi, K.
 Plunkett, *Rethinking Innateness: A Connectionist Perspective on
 Development*, The MIT Press, Cambridge, MA, 1996. 59
 [15] A. Fernald, Human maternal vocalizations to infants as biologically
 relevant signals: an evolutionary perspective, in: *The Adaptive Mind*,
 Oxford University Press, Oxford, 1992, pp. 391–428. 61
 [16] L. Gleitman, The structural sources of verb meanings, *Lang. Acquis.*
 1 (1) (1990) 1–55. 63
 [17] L. Gleitman, K. Cassidy, R. Nappa, A. Papafragou, J. Trueswell,
 Hard words, *Lang. Learn. Dev.* 1 (1) (2005) 23–64. 65
 [18] Z.M. Griffin, Why look? Reasons for eye movements related to
 language production, in: J. Henderson, F. Ferreira (Eds.), *The
 Integration of Language, Vision, and Action: Eye Movements and
 the Visual World*, Taylor & Francis, New York, 2004, pp. 213–247. 67
 [19] Z.M. Griffin, K. Bock, What the eyes say about speaking, *Psychol.*
Sci. 11 (2000) 274–279. 69
 [20] K. Hirsh-Pasek, R.M. Golinkoff, G. Hollich, An emergentist
 coalition model for word learning: mapping words to objects is a
 product of the interaction of multiple cues, in: *Becoming a Word
 Learner: a Debate on Lexical Acquisition*, Oxford Press, New York,
 2000, pp. 136–164. 73
 [21] P.W. Jusczyk, *The Discovery of Spoken Language*, The MIT Press,
 Cambridge, MA, 1997. 75
 [22] P. Kuhl, Early language acquisition: cracking the speech code, *Nat.*
Rev. Neurosci. 5 (2004) 831–843. 77
 [23] P.K. Kuhl, F.-M. Tsao, H.-M. Liu, Foreign-language experience in
 infancy: effects of short-term exposure and social interaction on
 phonetic learning, *Proc. Natl. Acad. Sci.* 100 (15) (2003) 9096–9101. 79
 [24] K.F. Lee, H.-W. Hon, R. Reddy, An overview of the sphinx speech
 recognition system, *IEEE Trans. Acoust. Speech Signal Process.* 38
 (1) (1990) 35–45. 81
 [25] P. Li, I. Farkas, B. MacWhinney, Early lexical development in a self-
 organizing neural networks, *Neural Networks* 17 (2004) 1345–1362. 83
 [26] B. MacWhinney, Competition and lexical categorization, in: F.
 Eckman, M. Noonan (Eds.), *Linguistic Categorization*, Benjamins,
 New York, 1989, pp. 195–242. 85
 [27] M. Maratsos, M. Chalkley, The internal language of children's
 syntax: the ontogenesis and representation of syntactic categories, in:
 K. Nelson (Ed.), *Children's Language*, vol. 2, Gardner Press, New
 York, 1980, pp. 127–214. 87
 [28] A.S. Meyer, A.M. Sleiderink, W.J. Levelt, Viewing and naming
 objects: eye movements during noun phrase production, *Cognition* 66
 (1998) B25–B33. 89
 [29] T.H. Mintz, E.L. Newport, T.G. Bever, The distributional structure
 of grammatical categories in speech to young children, *Cognit. Sci.* 26
 (2002) 393–424. 91
 [30] E.L. Newport, R.N. Aslin, Learning at a distance: I. Statistical
 learning of non-adjacent dependencies, *Cognit. Psychol.* 48 (2004)
 127–162. 93
 [31] M. Pena, L.L. Bonatti, M. Nespor, J. Mehler, Signal-driven
 computations in speech processing, *Science* 298 (5593) (2002)
 604–607. 95
 [32] S. Pinker, *Learnability and Cognition*, The MIT Press, Cambridge,
 MA, 1989. 97
 [33] K. Plunkett, Theories of early language acquisition, *Trends Cognit.*
Sci. 1 (1997) 146–153. 99
 [34] K. Plunkett, C. Sinha, M. Miller, O. Strandsby, Symbol grounding or
 the emergence of symbols? Vocabulary growth in children and a
 connectionist net, *Connect. Sci.* 4 (1992) 293–312. 101
 [35] W. Quine, *Word and Object*, The MIT Press, Cambridge, MA, 1960. 103
 [36] M. Redington, N. Chater, S. Finch, Distributional information: a
 powerful cue for acquiring syntactic categories, *Cognit. Sci.* 22 (4)
 (1998) 425–469. 105
 [37] T. Regier, *The Human Semantic Potential: Spatial Language and
 Constrained Connectionism*, The MIT Press, Cambridge, MA, 1996. 107
 [38] T. Regier, Emergent constraints on word-learning: a computational
 review, *Trends Cognit. Sci.* 7 (2003) 263–268. 109
 113

- [39] D. Richards, J. Goldfarb, The episodic memory model of conceptual development: an integrative viewpoint, *Cognit. Dev.* 1 (1986) 183–219.
- [40] D. Roy, A. Pentland, Learning words from sights and sounds: a computational model, *Cognit. Sci.* 26 (1) (2002) 113–146.
- [41] J.R. Saffran, Statistical language learning: mechanisms and constraints, *Curr. Dir. Psychol. Sci.* 12 (2003) 110–114.
- [42] J.R. Saffran, E. Johnson, R. Aslin, E. Newport, Statistical learning of tone sequences by human infants and adults, *Cognition* 70 (1999) 27–52.
- [43] J.R. Saffran, E.L. Newport, R.N. Aslin, Word segmentation: the role of distributional cues, *J. Mem. Lang.* 35 (1996) 606–621.
- [44] J.M. Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, *Cognition* 61 (1996) 39–61.
- [45] L. Smith, How to learn words: an associative crane, in: R. Golinkoff, K. Hirsh-Pasek (Eds.), *Breaking the Word Learning Barrier*, Oxford University Press, Oxford, 2000, pp. 51–80.
- [46] J. Snedeker, J. Trueswell, Using prosody to avoid ambiguity: effects of speaker awareness and referential context, *J. Mem. Lang.* 48 (2003) 103–130.
- [47] M.K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, J. Sedivy, Integration of visual and linguistic information in spoken language comprehension, *Science* 268 (1995) 1632–1634.
- [48] J. Tenenbaum, F. Xu, Word learning as bayesian inference, in: L. Gleitman, A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of Cognitive Science Society*, Erlbaum, Mahwah, NJ, 2000, pp. 517–522.
- [49] M. Tomasello, Perceiving intentions and learning words in the second year of life, in: M. Bowerman, S. Levinson (Eds.), *Language Acquisition and Conceptual Development*, Cambridge University Press, Cambridge, 2000, pp. 111–128.
- [50] M. Tomasello, N. Akhtar, Two-year-olds use pragmatic cues to differentiate reference to objects and actions, *Cognit. Dev.* 10 (1995) 201–224.
- [51] C. Yu, D.H. Ballard, R.N. Aslin, The role of embodied intention in early lexical acquisition, in: *Proceedings of the 25th Cognitive Science*

Society Annual Meetings, Erlbaum, Boston, MA, 2003, pp. 1293–1298. 31

[52] C. Yu, D.H. Ballard, R.N. Aslin, The role of embodied intention in early lexical acquisition, *Cognit. Sci.* 29 (6) (2005) 961–1005. 33



Chen Yu received his Ph.D. in Computer Science from the University of Rochester in 2004. He is an assistant professor in the Psychological and Brain Sciences Department at Indiana University. He is also a faculty member in the Cognitive Science Program and an adjunct member in the Computer Science Department. His research interests are interdisciplinary, ranging from human development and learning to machine intelligence and learning. He has received the 35 37

Marr Prize in the 2003 Annual Meeting of Cognitive Science Society. More information about Chen Yu's work can be found at <http://www.indiana.edu/~dll/>. 39 41 43 45



Dana Ballard is a professor of Computer Science and Brain and Cognitive Sciences at University of Rochester. His main research interest is in computational theories of the brain with emphasis on human vision. With Chris Brown, he led a team that designed and built a high-speed binocular camera control system capable of simulating human eye movements. The theoretical aspects of that system were summarized in a paper "Animate Vision", which received the Best 47 49 51 53 55

Paper Award at the 1989 International Joint Conference on Artificial Intelligence. Currently, he is interested in pursuing this research by using model humans in virtual reality environments. In addition he is also interested in models of the brain that relate to detailed neural codes. 57 59