Modeling disambiguation in word learning via multiple probabilistic constraints

Molly Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Young children tend to map novel words to novel objects even in the presence of familiar competitors, a finding that has been dubbed the "disambiguation" effect. Theoretical accounts of this effect have debated whether it is due to initial constraints on children's lexicons (e.g. a principle of mutual exclusivity) or situation-specific pragmatic inferences. We suggest that both could be true. We present a hierarchical Bayesian model that implements both situation-level and hierarchical inference, and show that both can in principle contribute to disambiguation inferences with different levels of strength depending on differences in the situation and language experience of the learner. We additionally present data testing a novel prediction of this probabilistic view of disambiguation.

Keywords: Word learning; mutual exclusivity; Bayesian models.

Introduction

A central property of language is that each word in the lexicon maps to a unique concept, and each concept maps to a unique word (Clark, 1987). Like other important regularities in language (e.g. grammatical categories), children cannot directly observe this general property. Instead, they must learn to use language in a way that is consistent with this generalization on the basis of evidence about only specific word-object pairs.

Even very young children behave in a way that is consistent with the one-to-one mapping between words and concepts in language. Evidence for this claim comes from what is known as the "disambiguation" effect. In a typical demonstration of this effect (e.g. Markman & Wachtel, 1988), children are presented with a novel and familiar object (e.g. a whisk and a ball), and are asked to identify the referent of a novel word ("show me the dax"). Children in this task tend to choose the novel object as the referent, behaving in a way that is consistent with the one-to-one word-concept regularity in language, across a wide range of ages and experimental paradigms (Mervis, Golinkoff, & Bertrand, 1994; Golinkoff, Mervis, Hirsh-Pasek, et al., 1994; Markman, Wasow, & Hansen, 2003; Halberda, 2003; Bion, Borovsky, & Fernald, 2013).

This effect has received much attention in the word learning literature because the ability to identify the meaning of a word in ambiguous contexts is, in essence, the core problem of word learning. That is, given any referential context, the meaning of a word is underdetermined (Quine, 1960), and the challenge for the world learner is to identify the referent of the word within this ambiguous context. Critically, the ability to infer that a novel word maps to a novel object makes the problem much easier to solve. For example, suppose a child hears the novel word "kumquat" while in the produce aisle of the grocery store. There are an infinite number of possible meanings of this word given this referential context, but the child's ability to correctly disambiguate would lead her to rule out all meanings for which she already had a name. With this restricted hypothesis space, the child is more likely to identify the correct referent than if all objects in the context were considered as possible referents.

What are the cognitive processes underlying this effect? There are broadly two proposals in the literature. Under one proposal, Markman and colleagues (1988; 2003) suggest that children have a constraint on the types of lexicons considered when learning the meaning of a new word — a "mutual exclusivity constraint." With this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Importantly, this constraint can be overcome in cases where it is incorrect (e.g. adjectives), but it nonetheless serves to restrict the set of lexicons initially entertained when learning the meaning of a novel word. Under this view, then, the disambiguation effect emerges from a constraint on the structure of lexicons.

Under a second proposal, the disambiguation effect is argued to result from online inferences made within the referential context (Clark, 1987; Diesendruck & Markson, 2001). Clark suggests that the disambiguation effect is due to two pragmatic assumptions held by speakers. The first assumption is that speakers within the same speech community use the same words to refer to the same objects ("Principle of Conventionality"). The second assumption is that different linguistic forms refer to different meanings ("Principle of Contrast"). In the disambiguation task described above, then, children might reason (implicitly) as follows: You used a word I've never heard before. Since, presumably we both call a ball "ball" and if you'd meant the ball you would have said "ball," this new word must refer to the new object. Thus, under this account, disambiguation emerges not from a higher-order constraint on the structure of lexicons, but instead from in-the-moment inferences using general pragmatic principles.

These two proposals have traditionally been viewed as competing explanations of the disambiguation effect. Research in this area has consequently focused on identifying empirical tests that can distinguish between these two theories. For example, Diesendruck and Markson (2001) compare performance on a disambiguation task when children are told a novel fact about an object relative to a novel referential label. They found that children disambiguated in both conditions and argued on grounds of parsimony that the same pragmatic mechanism was likely to be responsible for both inferences. More recent evidence contradicts this view: tests of children with autism, who are known to have impairments in pragmatic reasoning, find comparable performance on the disambiguation task between typically developing children and children with autism (Preissler & Carey, 2005; de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011). This result provides some evidence for the view that disambiguation is due to a domain-specific lexical constraint.

We suggest that this competing-alternatives approach to the disambiguation effect should be reconsidered. In a disambiguation task, learners may be making use of both higher-order knowledge about how the lexicon is structured as well as information about the pragmatic or inferential structure of the task. Both of these constraints would then support children's inferences. In other words, these two classes of theories may be describing distinct, but complementary mechanisms that each contribute to a single empirical phenomenon, with their weights in any given task determined by children's age and language experience, the nature of the pragmatic situation, and other task-specific factors.

The model described here explores this proposal computationally. We constructed a Bayesian model that captures effects of both inferences within individual situations and hierarchical inferences about the structure of lexicons. Inferences about individual situations are modeled using an intentional/pragmatic model of word learning (Frank, Goodman, & Tenenbaum, 2009), while generalizations about the nature of word-concept mappings are modeled as constraints on the set of lexicons that the model considers. We present a set of simulations and a developmental experiment showing that linguistic experience can influence the strength of disambiguation inferences at both levels.

The goal of our model is not to provide an algorithmic description of children's word learning, which we assume depends on psychological factors such as memory and cognitive control. Instead, we aim to provide an *ideal observer analysis*: to derive normative predictions given a well-articulated set of assumptions (Geisler, 2003). Human behavior can then be compared to this analysis, and deviations can be attributed to differences between the assumptions of the model and the realities of human psychology. Critically, neither our model (nor comparisons between it and human behavior) constitute claims of human optimality: Though our model employs optimal Bayesian inference, there is no implicit claim that human learners also do so (Frank, in press).

Model

We model a word learner as performing Bayesian inference to infer the structure of a lexicon l, which we represent as a (sparse) bipartite graph connecting words $W = w_1 ... w_n$ to objects $O = o_1 ... o_m$. We write the full possible set of lexicons as L. An example enumeration of such lexicons for the case of $n = m = 2$ is given in Fig. 2.

We assume a generative structure identical to the the model developed by Frank et al. (2009), with the added complexity of constraints placed on lexicons (described below; see Fig. 1). The critical feature of this existing model is that words are assumed to be generated by intentions. This feature allows the model to jointly solve the problems of mapping a word to an object in ambiguous contexts and learning a long term mapping between a word and concept.

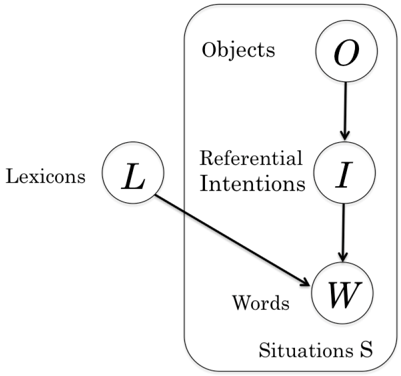

Figure 1: The generative process for our model.

The learner infers a distribution over lexicons, given a corpus S of situations (each consisting of sets of words $\bar{w}_s$ and objects $\bar{o}_s$). From Bayes' rule, the posterior probability of a lexicon is given by

$$P(l|S) = \frac{P(S|l)P(l)}{\sum_{l' \in L} P(S|l')P(l')} \quad (1)$$

We first define the likelihood term $P(S|L)$ and then return to the prior $P(L)$, which implements hierarchical constraints C on lexicons.

Using the generative process in Fig. 1, we can write the likelihood of a particular situation in terms of the relationship between the objects that were observed in the situation s, the speaker's referential intention i_s (a choice to speak about one of the objects), and the referring word w_s.[1] As in our prior work, we assume that referential intentions are unobserved and sum across all possible intentions uniformly[2]:

$$P(s|l) = \sum_{i_s \in \bar{o}_s} P(w_s, o_s, i_s | l) \quad (2)$$

By the conditional independence of words and objects, we can expand to:

$$P(s|l) = \sum_{i_s \in \bar{o}_s} P(w_s | i_s, l) P(i_s | o_s) \quad (3)$$

Finally, we aggregate across situations by taking the product of each independent situation:

[1] In this analysis, we focus only on situations where a single referring word is used.

[2] This assumption is made for purposes of simplicity only, as a variety of our previous work has explored the use of social and pragmatic information in biasing the distribution over intended referents (Frank & Goodman, 2012).

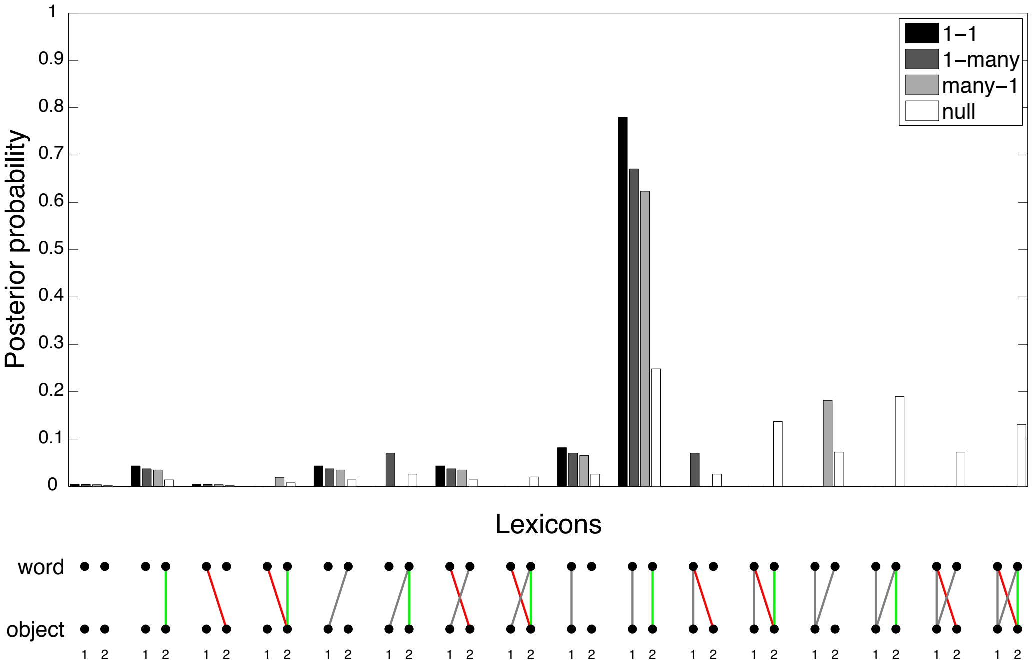

Figure 2: The posterior probability distribution over lexicons for our models for Simulation 1. Models were trained with situations establishing the mapping between w_1 and o_1 (the familiar word/object pair) and a disambiguation situation including w_2 and objects o_1 and o_2. The four different constraint models are distinguished by color in the main plot, while the 16 possible lexicons are shown on the horizontal axis. Lexicons are marked as links between words and objects, with the correct (w_2 and o_2) mapping marked in green and the incorrect (w_2 and o_1) mapping marked in red. The noise parameter α was chosen arbitrarily for display purposes and serves only to scale the results.

$$P(S|l) = \prod_{s \in S} \sum_{i \in \bar{o}_s} P(w_s|i_s,l)P(i_s|o_s) \tag{4}$$

We assume that there is some level of noise in both the choice of word given intention $P(w_s|i_s,l)$ and the choice of intention given object $P(i_s|o_s)$, such that the speaker could in principle have been mistaken about their referent or misspoken their word. We implement this decision by assuming a constant probability of random noise for each of these, which we notate α. For simplicity, α is assumed to be the same for both terms. The value of α serves only to scale the results we report below, but—as in nearly all probabilistic models—some level of uncertainty about the individual observations is necessary to make graded predictions.

We now consider the prior distribution over lexicons. We define this prior hierarchically as being the product of a constraint over lexicons $c \in C$:

$$P(l) = P(l|c)P(c) \tag{5}$$

We consider a hypothesis space of four different constraints placed on the mappings between words and objects within lexicons: one word to one object (*1-1* constraint), one word to many objects (*1-many* constraint), many words to one object (*many-1* constraint), and a null constraint. The 1-many constraint applies a restriction that each object maps to at most one word in a lexicon. The many-1 constraint applies a restriction that each word maps to at most one object in a lexicon. The 1-1 constraint applies both of these restrictions, and the null constraint applies neither of these restrictions.[3] In practice, these hypotheses were implemented such that each lexicon consistent with a constraint was equiprobable, and all inconsistent lexicons had probability 0. For simplicity, we assumed that $P(c) \propto 1$, although this assumption could easily be modified in future work.

For the simulations below, we were able to infer exact posterior distributions by enumerating all possible lexicons and normalizing (Equation 1).

Simulation 1: Disambiguation at multiple levels

As a first test of our model on the disambiguation task, we trained the model on a corpus containing two situations. The first was an unambiguous situation in which word w_1 was associated with object o_1. This piece of evidence corresponded to the known word in the disambiguation task ("ball" in the example described above). We also included a disambiguation experimental situation, where the previously learned object o_1 (ball), a new object o_2 (whisk), and a novel word w_2

[3] The 1-many constraint is related to the concept learning model proposed by Goodman, Tenenbaum, Feldman, and Griffiths (2010) using a disjunctive normal form grammar.

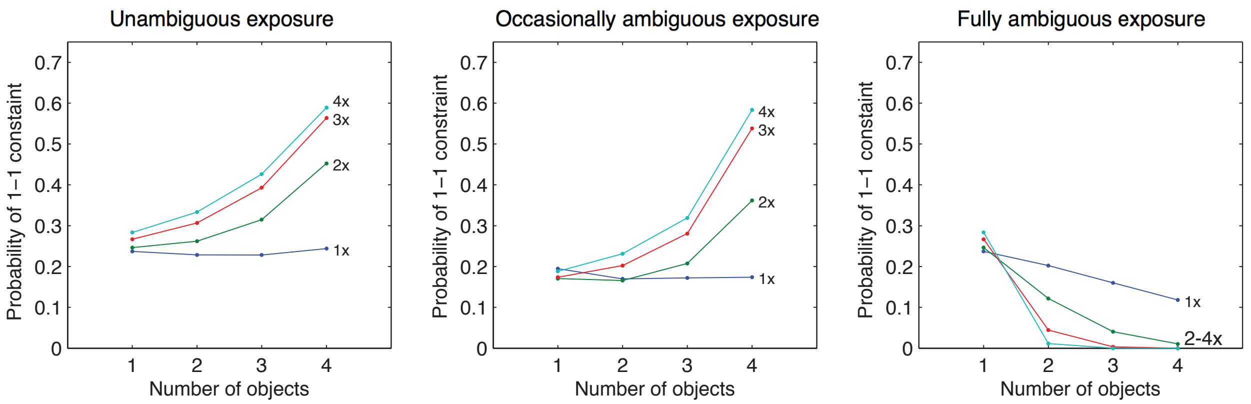

Figure 3: Results from Simulation 2. Each panel shows the probability of inferring a 1-1 constraint on the lexicon given a different input corpus, described in text. Horizontal axes varies the overall number of distinct objects presented in the exposure corpus, while the colored lines denote different numbers of exposures to the corpus.

("dax").

Figure 2 shows the posterior distribution over lexicons inferred on the basis of this corpus. Each of the 16 possible lexicons (assuming a world with only two words and two objects) are represented along the x-axis, where lexicons are represented by object and word nodes connected by links.

In this maximally simple simulation of the disambiguation task, all four prior constraints give the highest posterior probability to the lexicon that links the novel word and the novel object. This result emerges from the structure of the inference problem: Given that the learner has already observed an association between w_1 and o_1, lexicons that posit a link between w_1 and o_2 are less probable than those that posit a link between w_2 and o_2. This result comes about because an object with two names (w_1 and w_2) can be talked about in two different ways, and each of them is individually less probable than the one way of talking about an unambiguously-named object. (This result echoes the finding of mutual exclusivity in Frank et al., 2009).

These results suggest that disambiguation behavior in children could emerge without a 1-1 constraint on lexicons. On the other hand, prior constraints affected the *strength* of the disambiguation inference. Constraints barring 1-many and many-1 mappings increased the posterior probability of the correct lexicon; when both were in place, the correct lexicon had by far the highest probability. Thus, probabilistic inference and hierarchical constraints both support disambiguation behavior in the model.

Simulation 2: Learning constraints on lexicons

Simulation 1 suggested that a learner could behave consistent with a 1-1 constraint on lexicons without assuming a hard constraint on the structure of lexicons. Nevertheless, imposing such a hard constraint raised the probability of a correct answer on the disambiguation task. In this simulation, we show that learners may induce a higher-order constraint on lexicons given the right kind of evidence.

To explore the model's ability to learn a hierarchical 1-1 constraint on lexicons, we trained our model on three corpora. Each corpus consisted of a set of situations with a single word and a single object, but we varied whether these mappings were consistent. The first, the "unambiguous exposure" corpus, showed unambiguous (1-1) mappings between words and objects. The second, the "occasionally ambiguous" corpus, showed the same body of data but with two contradictory mappings appended to the end. The final, "fully ambiguous," corpus consisted of one word that mapped to many objects. We varied both the number of exposures to the corpus (1–4) and the number of objects in the corpus (1–4). We then examined the posterior probability of the 1-1 constraint given these exposure corpora.

The results of this simulation are shown in Figure 3. Given 1-1 evidence, the model induces a 1-1 constraint on lexicons, and this bias becomes stronger as the number of observations increases. The posterior probability of the 1-1 constraint is decreased only slightly by a few ambiguous observations, reflecting the general robustness of this inference. In contrast, in the fully ambiguous condition, the model learns with relatively little data that a 1-1 constraint does not hold.

Simulation 3: Stronger mappings result in stronger disambiguation

In Simulation 3, we explore whether providing more evidence for a link between the known word and object will in turn strengthen the probabilistic disambiguation effect between words and objects. Recall that the disambiguation effect in Simulation 1 emerged as a result of prior evidence for an association between the known word and object ("ball" and ball). Thus, this model predicts that if the learner receives more evidence for an association between the known word and known object, the disambiguation bias should become stronger.

We trained the model with either 1, 2, or 3 situations in which w_1 was unambiguously associated with o_1. We then tested the model in the disambiguation task with a known and

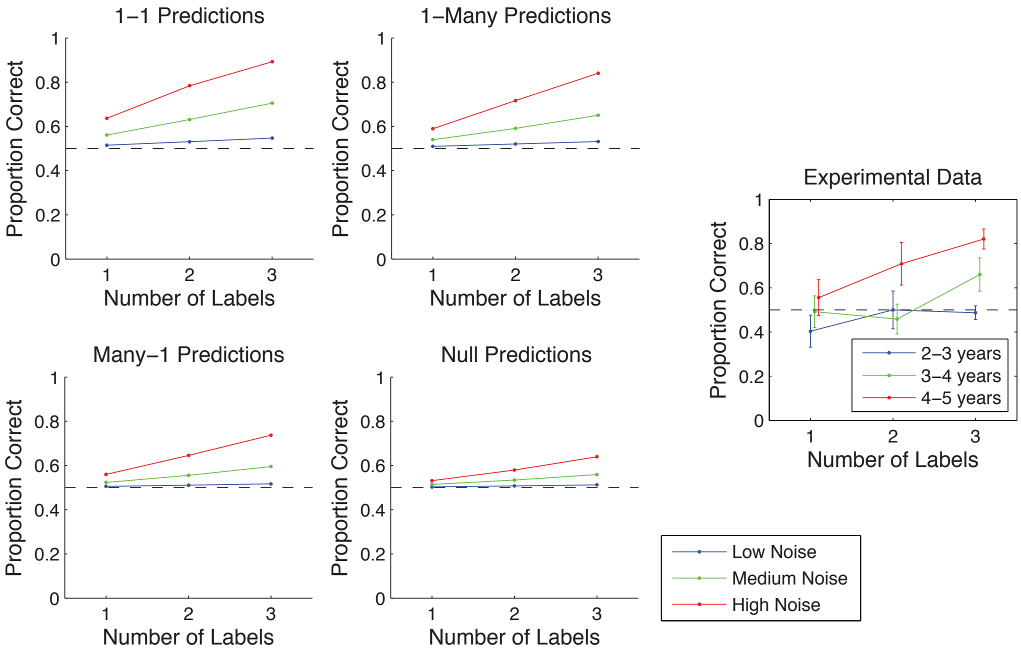

Figure 4: Model predictions under each of the four lexical constraints (left) and experimental results (right) for success in the disambiguation task as a function of the number of labels observed in training. Lower legend shows noise conditions for the model simulations.

unknown object and a novel word, as in Simulation 1, but using a Luce choice rule to compute the probability of a correct choice (Luce, 1963). If more observed associations between the known word and object lead to a stronger bias toward correct lexicons, we should expect the disambiguation bias to increase with the number of training situations.

Assuming a 1-1 constraint, the magnitude of the bias toward correct lexicons increases with number of training situations with the known word–known object association (Fig. 4, upper–left panel). In addition, the magnitude of this increase is sensitive to the noise parameter α that determines the probability that the wrong word was spoken to refer to an object.

Experiment

We tested the prediction that confidence in the known word mapping leads to a stronger disambiguation inference in preschool children.

Methods

We recruited 110 children ages 2;1–4;11 from the floor of the Boston Children's Museum. In each one-year age group, we collected data from 35–38 children.

Each child completed four trials. Each trial consisted of a training and a test phase in a "novel-novel" disambiguation task (de Marchena et al., 2011). In the training phase, the experimenter presented the child with a novel object, and explicitly labeled the object with a novel label 1, 2, or 3 times ("Look at the dax"), and contrasted it with a second novel object ("And this one is cool too") to ensure equal familiarity. In the test phase, the child was asked to point to the object referred to by a second novel label ("Can you show me the zot?"). Number of labels used in the training phase was manipulated between subjects. There were eight different novel words and objects. Object presentation side, object, and word were counterbalanced across children.

Results

Responses were coded as correct if participants selected the novel object at test. As predicted, children showed a stronger disambiguation effect as the number of training labels increased, and as noise decreased with age (Fig. 4, right panel).

We analyzed the results using a logit mixed model to predict correct responses with age and number of labels as fixed effects, and participant as a random effect. There was a significant effect of age ($\beta = .044$, $p < .001$) such that older children showed a stronger disambiguation bias. There was also a significant effect of number of labels, such that more training labels led to stronger disambiguation ($\beta = .454$, $p < .001$). The interaction between age and number of labels was not significant ($\beta = .019$, $p = .16$). Children's increased confidence in the disambiguation inference, as a function of number of training labels, is consistent with model predictions.

General Discussion

The disambiguation effect suggests the presence of underlying cognitive mechanisms that help children solve the difficult mapping problem inherent of early word learning (Quine, 1960). Two classes of mechanisms have been proposed: a constraint on the structure of permitted lexicons, and in-the-moment pragmatic inferences about the most likely referent given the context. We used a hierarchical Bayesian model to explore the independent contributions of these two effects and find that neither mechanism is necessary to create a bias, but either is sufficient. Disambiguation is strongest when both mechanisms jointly contribute.

This result has important consequences for attempts to experimentally differentiate between the proposed accounts of disambiguation. Given that both mechanisms can in principle lead to disambiguation behavior, experimental tests of disambiguation cannot distinguish between these two theories (as they are instantiated here). That is, evidence for disambiguation behavior is consistent with both a pragmatic account and a mutual exclusivity constraint account. Furthermore, there may be variability in the weights of these constraints across populations. For example, higher-order lexical constraints may play a larger role in disambiguation for individuals with impaired social-cognitive skills (e.g. autism), relative to typically developing children. Our results suggest that future research in this area should reconsider the assumption that a single mechanism must completely and independently give rise to the disambiguation effect.

Our model may provide useful insight into disambiguation in bilingualism. For bilingual learners, the structure of associations between words and objects in the environment differs from that of monolinguals. Bilingual learners typically observe two basic-level words associated with each object rather than one. To make sense of these associations, they might ultimately form an overhypothesis that there is a 1-1 constraint on lexicons within each language, but they might nevertheless initially entertain a 1-many constraint as a hypothesis. Indeed, there is evidence that disambiguation behavior is weaker in bilingual and trilingual children (Byers-Heinlein & Werker, 2009).

Finally, it is important to consider the limits of an ideal observer analysis. While our results suggest that both mechanisms could contribute to disambiguation behavior, this finding does not entail that both mechanisms do in fact contribute. It remains possible that disambiguation behavior is the result of a single mechanism. Nonetheless, given evidence from other domains that the mind may simultaneously integrate basic probabilistic inferences with higher-order constraints (Tenenbaum, Kemp, Griffiths, & Goodman, 2011), it seems likely that disambiguation behavior emerges from multiple underlying cognitive mechanisms.

Acknowledgements

We gratefully acknowledge Jesse Snedeker for the suggestion that led to the Experiment, as well as for providing materials from de Marchena et al. (2011). We thank Joseph Austerweil for valuable feedback, and Melina Flores for assistance in data collection.

References

Bion, R., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*, 39–53.

Byers-Heinlein, K., & Werker, J. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*, 815–823.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of Language Acquisition. Hillsdale, NJ: Erlbaum.*

de Marchena, A., Eigsti, I., Worek, A., Ono, K., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, *119*, 96–113.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*, 630.

Frank, M. C. (in press). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*.

Frank, M. C., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578.

Geisler, W. (2003). Ideal observer analysis. *The Visual Neurosciences*, 825–837.

Golinkoff, R., Mervis, C., Hirsh-Pasek, K., et al. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*, 125–125.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2010). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*, B23–B34.

Luce, R. (1963). Detection and recognition. *Handbook of Mathematical Psychology*, *1*, 103–189.

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

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*, 241–275.

Mervis, C., Golinkoff, R., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, *65*, 1163–1177.

Preissler, M., & Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. *Cognition*, *97*, B13–B23.

Quine, W. (1960). *Word and object* (Vol. 4). The MIT Press.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.