Towards semantically rich and recursive word learning models

Francis Mollica (fmollica@bcs.rochester.edu)
Steven T. Piantadosi (spiantadosi@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA

Abstract

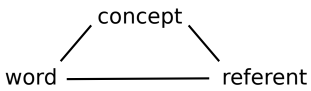
Current models of word learning focus on the mapping between words and their referents and remain mute with regard to conceptual representation. We develop a cross-situational model of word learning that captures word-concept mapping by jointly inferring the referents and underlying concepts for each word. We also develop a variant of our model that incorporates recursion, which entertains the idea that children can use learned words to aid future learning. We demonstrate both models' ability to learn kinship terms and show that adding recursion into the model speeds acquisition.

Keywords: word learning; cross-situational learning; language acquisition

Introduction

Most contemporary research on word learning examines how children solve the word-to-referent *mapping problem*—i.e., how children who are presented with multi-word utterances in multi-referent contexts learn which objects, events, or properties a word refers to. Real-life word learning is actually much more interesting than discovering these simple correspondences; discovery of the meaning of words often goes hand-in-hand with true conceptual learning. Waxman and Markow (1995) argue that, for kids, encountering a new word provides an "invitation" for a conceptual distinction to be made. It has even been shown in adults that having novel words improves category learning (Lupyan, 2006). In this light, it is not surprising that having conceptual knowledge influences how children extend novel words (Booth & Waxman, 2002).

Here, we aim to extend formalized theories of cross-situational word learning to capture not just the salient physical referents that are present, but the more abstract conceptual meanings that adults posses. A useful framework for understanding our approach is the semiotic triangle put forth by Peirce (1868):



Competent word users must know each link—the mappings from words to referents and words to abstract concepts, and the relationship between abstract concepts and their referents. In learning, children must use observed co-occurrences of, say, the word "accordion" and physical accordions to infer this mapping, as well as the abstract concept ACCORDION that can, in principle, refer to infinitely many physical objects.

Cross-situational models of word learning capitalize on the fact that a word is often heard when its referent is in the immediate context. Repeated exposure of words and their corresponding referents in multiple contexts provides the basis for the statistical learning of the association between words and referents. Variants of cross-situational models have been couched in terms of connectionist networks (Plunkett, Sinha, Møller, & Strandsby, 1992), deductive hypothesis testing (Siskind, 1996), hypothesis competition (Regier, 2005) and probabilistic inference (Yu & Ballard, 2007; Frank, Goodman, & Tenenbaum, 2009). These prior theories remain mute about concept-word mappings or assume words label preexisting concepts, an assumption that is inherently problematic. These theories leave unanswered both how conceptual representations develop and how the word-concept mapping interacts with the word-referent mapping.

An alternative approach is to treat conceptual development and word learning as a joint inference problem. This is the approach that we develop in this paper building on general versions of ideas proposed in prior cross-situational learning models. Our work provides a cross-situational word learning model that aims to integrate the three aspects of word learning: word labels, object referents, and abstract concepts. It works by combining a plausible (though simplified) semantic representation with cross-situational evidence of words and referents. Our model also captures the idea that children do not just learn concepts but organize conceptual information to form intuitive theories about the world (Carey, 1985; Wellman & Gelman, 1992). There have been advances in computational modeling on learning how these theories might be structured (Tenenbaum, Griffiths, & Kemp, 2006) and acquired (Ullman, Goodman, & Tenenbaum, 2012).

Intuitive theories likely influence how a child approaches word learning. For example, a child that has learned an abstract conceptual structure might approach learning words as denoting relationships across the structure; this kind of behavior is likely relevant in number word learning, where the structure of the count list ("one", "two", "three", etc.) likely provides a *placeholder structure* for their meanings (Carey, 2009). Of course, such structures can be more complex than lists: a child's theory of kinship as a family tree might shape how they approach the task of learning kinship terms. For instance, a child might need to have the right conceptual structure (e.g. a tree) with the referents in their particular family members in the right place in order to correctly determine abstract relations like *uncle*.

Our Approach

Here, we study the domain of kinship as an example of cross-situational word learning that requires abstract conceptual knowledge. Kinship is an ideal domain because it lends itself to straightforward logical representation and it is one of the early domains available to children, building on their initial learning of terms like *mom* and *dad*. The domain is complex enough to need interesting cognitive mechanics, but simple enough to be computationally and representationally tractable (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008). In our demonstration, we focus on learning the word-concept-referent mappings for following kinship terms of English: *parent, child, spouse, grandparent, sibling, uncle/aunt*[1] and *cousin.*

We start by considering a cross-situational learning setup, meaning that children and the model observe both words and immediately available referents (e.g., *parent* spoken by "Rose" to refer to "Brandy"). The model formalizes a semantic space that includes the possibility of learning individual referents for each word (as in traditional word-learning models), or more abstract logical concepts. This representation can be thought of as a function that, given a context, returns a set of referents in that context. The simplest hypotheses explicitly "memorize" the set of referents for each word; however, the model also allows logical hypotheses that implicitly define this set. For instance, a word like *parent* might return the pairs X,Y such that X is the parent of Y. The model is cross-situational because any instance of *parent* will occur with only one particular X and Y (e.g. "Brandy" and "Rose"). The learner must aggregate information across usages in order to both figure out the more abstract, productive form of the meaning, and learn that *parent* does not refer to a particular parent (e.g. "Brandy").

Our learning model uses two components, both of which have been used in previous models of conceptual and language learning (e.g. Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Ullman et al., 2012; Piantadosi, Tenenbaum, & Goodman, 2013, 2012): a simplicity prior over semantic representations and a size principle likelihood specifying how well any hypothesized representation explains the observed data.

Our prior takes the form of a Probabilistic Context Free Grammar (PCFG) which specifies how learners may combine our assumed semantic primitives and entities in the context (see Table 1).

Table 1: The PCFG used for learning Kinship terms.

START → SET	SET → parents(SET)
SET → union(SET,SET)	SET → children(SET)
SET → intersection(SET,SET)	SET → spouses(SET)
SET → set_difference(SET,SET)	SET → male(SET)
SET → complement(SET)	SET → female(SET)
SET → *specific referent*	SET → X

Our PCFG for learning kinship terms included the set-theoretical primitives, union, intersection, set-difference and complement, and primitives specific to the kinship domain, parents, children, spouses, male and female. All entities in the context were potential sets. Additionally, the speaker X was included in the grammar as a potential set. The context for our kinship model was based on the family tree shown in Figure 1. All members of the family tree were seen by the model as potential referents. We assume the learner has developed the abstract structure of a family tree, including the primitive relations between entities. Future research will attempt to integrate learning the tree structure and primitives into the model.

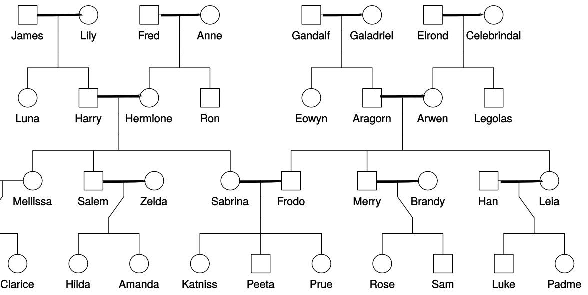


Figure 1: Family tree serving as the context for our model. Bold lines signify spouse relationships.

The likelihood function gives the probability of a word W_i correctly being mapped onto a referent Y_i conditioned on the speaker X_i, the context C_i, and the current hypotheses for each word, i.e. the hypothesized lexicon L. We assume a noisy likelihood process, where a correct word-referent pair is observed with probability α and an incorrect word-referent pair is observed with probability 1-α. Here, we fix $\alpha = 0.9$.

The PCFG and the likelihood function specify a probability model for all possible lexicons. With this model we can rank the probability of a hypothesized lexicon conditioned on observed word-referent mappings in a given context with a given speaker according to Bayes Rule:

$$P(L|W,X,Y,C) \propto P(L) \cdot \prod_i P(W_i,Y_i|L,X_i,C_i) \qquad (1)$$

Here, $P(L)$ is the probability of L under the PCFG and $P(W_i,Y_i|L,X_i,C_i)$ gives the likelihood of the word-referent mappings under the hypothesized lexicon and the observed data. The PCFG prior penalized *complex* lexicons, meaning that this builds in a simplicity bias, a natural assumption for learners (Feldman, 2003) especially for the kinship domain, where it has been shown that kinship systems are the optimal trade-off between simplicity and informativity (Kemp & Regier, 2012). Thus, learners "score" any hypothesized lexicon (mapping of words to meanings) L by considering (i) how complex L is and (ii) how well L explains the observed word-referent usages.

Methods

Using Equation 1 to determine the most likely lexicons given the data is a complex inference problem because there are, in principle, infinite possible lexicons generated from the PCFG. Here, we solve the problem using sampling—Markov-Chain Monte-Carlo (MCMC)—methods. MCMC provide samples from the posterior distribution (in this case $P(L|W,X,Y,C)$)

[1] For simplicity, we do not distinguish gender here, although there is nothing to suggest the model could not handle it with the addition of gender primitives.

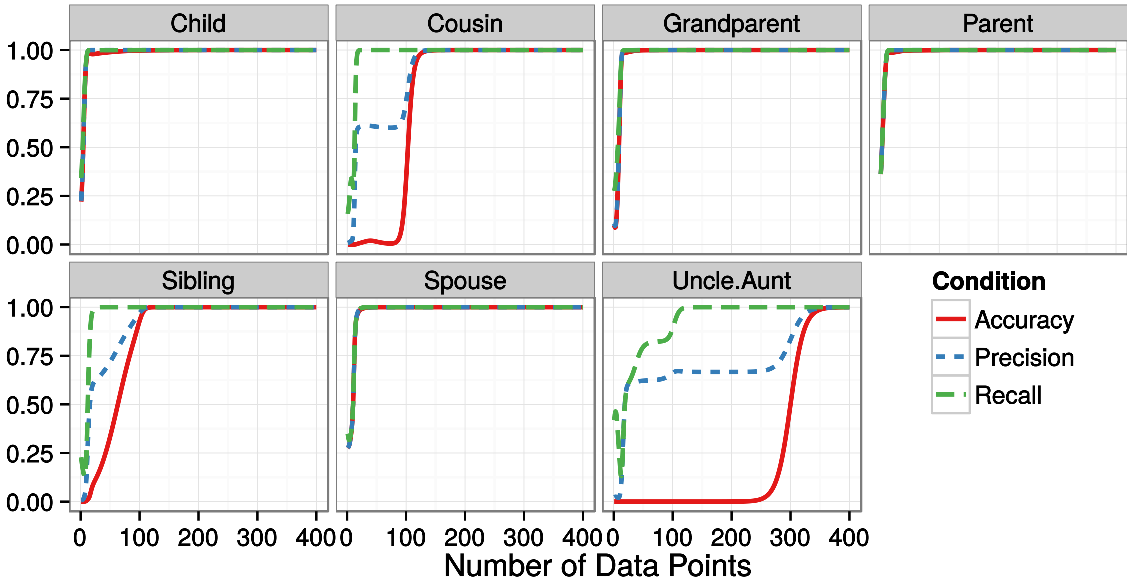


Figure 2: Expected proportion of corrected hypotheses by the number of observed data points.

by walking around the space of hypotheses. In the limit, these walks provably draw samples from the true posterior distribution. Because our hypothesis space is discrete, this method primarily allows us to determine the most likely lexicon given the observed data.

We ran the model on simulated data constructed by sampling data from a correct lexicon with the specified noise parameter α. Note that by sampling based on the true word-referent mappings in the tree, the data selected is biased towards the more common relationships in the specific tree used as context. Additionally, we did not fix the speaker to a specific person. If we had fixed the speaker, the data would be egocentric with regard to the speaker, meaning that the model would be attempting to learn the referents of kinship terms for that person as opposed to the underlying concepts for kinship terms in general.

We varied the amount of data the model received from 10 data points to 250 data points at ten point intervals and ran 16 sampling chains for one million steps at each data amount. For each chain, we saved the 100 lexicons with the highest posterior score, and used the union of these sets as a finite hypothesis space representing the posterior distribution (Piantadosi et al., 2012). We used the finite hypothesis space to calculate the learning trajectory for each word as the amount of data observed increases. Given that the hypotheses were generated based on varying amounts of data, their posterior score and likelihood per data point were normalized by recalculating them on a set of 1000 data points. The growth trajectory is then represented as the posterior-weighted average of all lexicons' accuracy.

Model Results

The model learned the correct hypothesis for each word[2] (see Table 2). As can be seen in Figure 2, the model learns the correct hypotheses for simple concepts, such as PARENT and GRANDPARENT, faster than it learns the correct hypotheses for more complex concepts, such as SIBLING and COUSIN. The logistic shape of the growth curves suggests that accurate performance is not gradual. Therefore, the model predicts that after observing a certain amount of data, a child should be able to learn the correct concept for each word.

The posterior-weighted average of each hypothesis' recall and precision[3] can be used to identify patterns of over-generalization and under-generalization of a word's referents. The posterior-weighted average recall represents the proposed hypotheses ability to select the correct referents. A recall of one means that on average each hypothesis selected all of the correct referents. The posterior weighted average precision reflects the hypotheses' specificity in selecting only the correct referents. Recall greater than precision suggests that a word is being over-extended to incorrect referents as can be seen for *siblings, cousins* and *uncles/aunts*.

As expected by the pattern of precision and recall, the incorrect hypotheses in the finite hypothesis space (see Table 2) tend to over-generalize terms corresponding to complex concepts. The most common incorrect hypotheses for *siblings, uncles/aunts* and *cousins* are over-extensions of the terms to respectively include the speaker herself, the speaker's parents, or everyone in the speakers generation including the speaker.

[2] For readability, all hypotheses presented in the paper have been transformed into their simplest semantically equivalent form—i.e., the shortest composition of primitives denoting the same set of referents.

[3] Recall is the amount of correct referents proposed by the hypothesis divided by the total amount of correct referents for the word. Precision is the total amount of correct referents proposed by the hypothesis divided by all referents proposed by the hypothesis.

Table 2: Correct hypothesis and most common **incorrect** hypothesis for each word.

Word	Correct Hypothesis	Most Common Incorrect Hypothesis
Children	children(X)	children(spouses(X))
Parents	parents(X)	spouses(parents(X))
Grandparents	parents(parents(X))	parents(spouses(parents(X)))
Spouses	spouses(X)	male(spouses(X))
Siblings	set_difference(children(parents(X)), X)	children(parents(X))
Uncles/Aunts	union(set_difference(children(parents(parents(X))), parents(X)), spouses(set_difference(children(parents(parents(X))), parents(X))))	union(children(parents(parents(X))), spouses(children(parents(parents(X)))))
Cousins	children(union(set_difference(children(parents(parents(X))), parents(X)), spouses(set_difference(children(parents(parents(X))), parents(X)))))	children(spouses(children(parents(parents(X)))))

Interestingly, the most common mistakes made learning the simpler concepts tend to under-generalize a term by imposing an additional constraint. For *children, parent* and *grandparent*, the incorrect hypotheses would be correct if every child had two married parents, represented by spouse relationships. The incorrect hypothesis for *spouse* places an unnecessary male constraint on the correct hypothesis. At face value, these mistakes seem like plausible mistakes a child learning kinship terms could make.

The model also demonstrated that under simple assumptions, rational statistical learners will not learn specific referents but prefer abstract logical hypotheses. For instance, the posterior probability of a hypothesis containing any specific referent was at most 10^{-25} by 5 data points. The pressure for abstraction occurs because any particular referent is unlikely in the prior and the speaker varies, meaning that accurate word-referent maps are hard to create[4]. For example, proposing that *sibling* refers to *rose* is only true when spoken by Rose's siblings. To avoid under-generalizing, the model constructed abstract hypotheses that over-generalized the word to many referents. As the model observed more data, it narrowed down the set of potential referents in an attempt to find the simplest correct hypothesis.

Comparison to Child Data Our model's predicted learning trajectories can, in principle, be compared to child acquisition data. We used the Words and Gestures MacArthur-Bates Communicative Development Inventory (MCDI) data from Word Bank[5] to calculate child learning trajectories for *parents, grandparents, siblings* and *uncle/aunts*. The MCDI data is a parent report measure of their child's understanding of a specific list of words. As our model did not differentiate gender, we averaged over the gendered terms in the MCDI for each concept in our model (e.g., *mommy/daddy* corresponds to *parents*). While the MCDI is the best data set available, the child's understanding of a word is based on parental report rather than experiments and, thus, might incorrectly capture children's true understanding. For example, without proper controls, over-generalization can be mistaken for correct understanding.

Figure 3a shows smoothed growth curves fit to MCDI data. The order of acquisition that the model predicts roughly matches the qualitative pattern observed in this data: *mommy/daddy* (*parent*) is learned quickly, *grandma/grandpa* (*grandparent*) is learned somewhat less quickly, and *brother/sister* (*sibling*) and *uncle/aunt* take much more time. Intuitively, the model explains this acquisition trajectory by penalizing complex hypotheses. When the model has not observed much data, it relies heavily on its prior, which is biased to favor simplicity. If you compare the complexity of the correct hypotheses the model learns (see Table 2), PARENT is a single semantic primitive, GRANDPARENT requires two primitives, and SIBLINGS and UNCLES/AUNTS are much more complex.

However, beyond the relative difficulty of words, the general shape of the model predictions do not closely match children's trajectory. The child data suggest a gradual acquisition of kinship terms. As discussed in Ullman et al. (2012), one possible explanation for this discrepancy is that the model predicts an individual child's learning trajectory; whereas, the MCDI data is an average growth trajectory over children. If we consider that children might differ in the rate at which they observe data, curves that are individually logistic might suggest gradual learning when averaged together.

To further explore the relationship between children's behavior and the model, we considered transforming both learning curves to relate performance to the number of instances of each word; however, there is no directly analogous way to convert the child data. We tried fitting a logistic regression for each word as a function of age in months and dividing the coefficient for age by an estimate of the number of instances of that word a child hears in a month. In doing so, we assumed children hear 360,000 words per month and estimated instances of a specific word using frequency data from CHILDES[6] (see Figure 3b).

Interestingly, this transformation suggests that higher frequency words need **more** data to be learned. For instance, *mommy*, which is learned very quickly, is also extremely frequent. Its learning curve, therefore, stretches out, showing that children require many instances to learn. Conversely, *uncle/aunt*, which is learned slowly, is relatively infrequent and its learning curve suggests that children require very few instances to learn.

However, this pattern in the transformed data may not reflect what happens with children. The word frequency estimates from corpus data may overestimate the amount of

[4] When the model is provided with ego-centric data, hypotheses with specific referents are more likely.

[5] Retrieved from wordbank.stanford.edu on 2015-01-20.

[6] Data retrieved on 2015-01-20 using ChildFreq (Bååth, 2010).

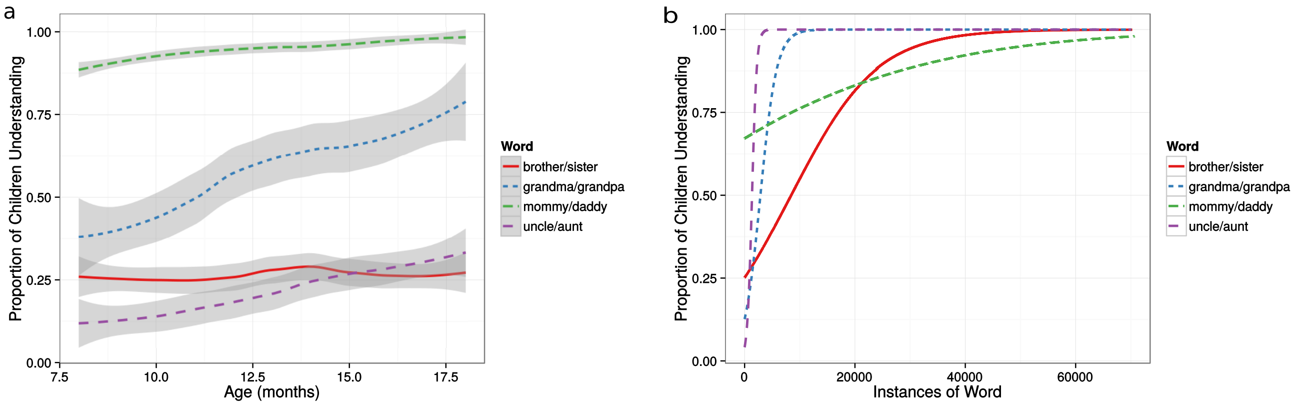


Figure 3: (*a*) Growth trajectories of MCDI data by age. (*b*) Growth trajectories from MCDI data by number of instances.

data actually considered by the learner. At the same time, if the pattern of learning trajectories from the transformed child data is accurate, there is a discrepancy with the model. According to the child data, the more complex concepts such as UNCLE/AUNT require relatively few data points and the simple concepts require many data points. This is the exact opposite prediction from our model and any other with a simplicity bias. One possibility for this discrepancy is that while word learning is a joint learning problem, the learning of concept-word mappings and word-referent mappings might not occur on the exact same time scale. In this domain, children might learn word-referent mappings before they learn word-concept mappings. As a result, the transition from associative link to abstract conceptual representation might require additional data points. Whether or not word-referent mappings must precede word-concept mappings is an empirical question.

Another possible explanation is that in kinship and other semantic domains, a conceptual representation's dependence on an abstract structural representation (e.g., tree or taxonomy) limits a child's ability to learn word-concept mappings until the abstract structural representation has been developed. This explanation could plausibly explain the transformation requiring more words for simple concepts. If you consider the learner using early relationships, such as PARENT and SPOUSE, to construct a family tree, the high instance requirement might reflect the development of the kinship tree. Future research will explore these possibilities.

Recursive vs. Non-Recursive

One of the most interesting aspects of children's theory learning is that theories can be rich, interconnected systems of ideas and concepts. Our model allows us to explore learning interconnected representations in the domain of kinship by potentially allowing words to be defined in terms of other words, a capacity essentially for *recursion*.

In a second version of the model, we built recursive rules into the PCFG. For instance, the recursive grammar allows a function like SET → uncle(SET), giving the referents that are considered "uncles" by the model's current hypothesized lexicon (which may or may not be correct). To prevent infinite recursion, hypotheses were restricted to recurse maximally ten times.

As with the non-recursive model, we varied the amount of data given to the model from 5 data points to 200 data points at five point intervals and ran 16 sampling chains for 500,000 steps at each data amount. Again, we created a finite hypothesis space and calculated the learning trajectories using the same method as before (see Figure 4).

For the words with simple hypotheses, there is no substantial difference in allowing the model to recurse; however, for the more complex hypotheses, allowing recursion decreases the number of data points that need to be observed. This is because some word meanings can be expressed more concisely by referencing *other* word meanings. For instance, COUSIN becomes easier to learn once UNCLE is known because COUSIN can be expressed as children(uncle(X)), instead of the more arduous form in Table 2 above.

As our model shows, acquisition of this kind of recursive, interrelated theories is possible through essentially the same mechanisms as non-recursive theories. By creating a representation language that permits recursion and doing inference over that language with cross-situational data, we are able to learn word meanings that are richly interconnected. A general prediction of this system is that permitting recursion of the referents of concepts to each other will speed learning of certain types of meanings by allowing them a much more concise representation. This exact mechanism might also account for how complex semantic primitives develop from simple primitives.

Discussion

In this paper, we have provided a possible mechanism for simultaneously learning a conceptual representation and the mapping of that representation to a word. This model differs from previous models attempting to learn word-concept mappings (e.g. Fazly, Alishahi, & Stevenson, 2010) in that we focus on the learning of concepts requiring abstract relations between entities. We offer this model as a first step in suggesting that children could learn hierarchical concepts and their corresponding words jointly. We expect that this mechanism will generalize to other hierarchical domains.

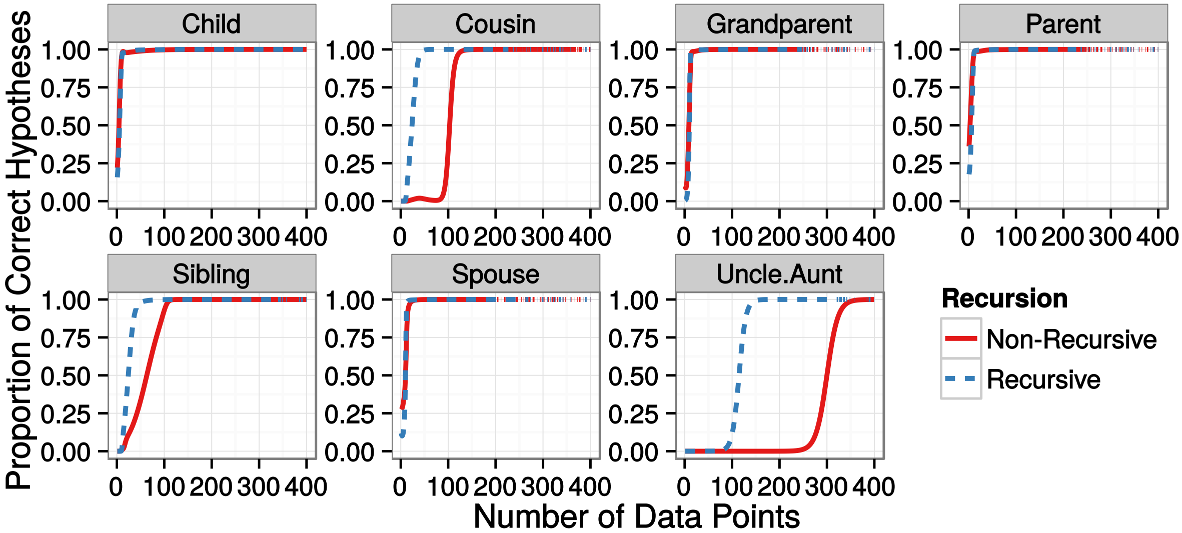


Figure 4: Growth trajectories of the model with and without recursion.

Here, we focused on conceptual development as the formulation and refinement of hypotheses about the relationship between speakers and referents on a pre-existing kinship tree. Future work will attempt to integrate the tree construction into the model.

Both our model and other models of cross-situational concept-word mapping learning rely heavily on the semantic primitives or features posited to represent concepts. The scalability of our model depends on uncovering what primitives people use when constructing conceptual representations. Ongoing research is focused on discovering the primitives people use and future research will investigate both the development of complex primitives from simple primitives and the time course of primitive development.

Conclusion

We have developed a cross-situational word learning model that captures richer semantic representations than associative links. Instead, it captures the acquisition of abstract semantic relations in the context of a rich theory of the world. We developed two variants of the model: one which permitted recursion and one which did not. We show that the recursive model not only can be made to work—explaining how word learning may interface with rich semantic theories—but also speeds acquisition.

References

Bååth, R. (2010). Childfreq: An online tool to explore word frequencies in child language.

Booth, A. E., & Waxman, S. R. (2002). Word learning is smart: Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, *84*(1), B11–B22.

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Lupyan, G. (2006). Labels facilitate learning of novel categories. In *The sixth international conference on the evolution of language* (pp. 190–197).

Peirce, C. S. (1868). On a new list of categories. In *Proceedings of the american academy of arts and sciences* (Vol. 7, pp. 287–298).

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2013). Modeling the acquisition of quantifier semantics: a case study in function word learnability. *Under review*.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3-4), 293–312.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.