Joint acquisition of word order and word reference

Luke Maurits (luke.maurits@adelaide.edu.au)
Amy F. Perfors (amy.perfors@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide
Adelaide SA 5005, Australia

Abstract

Inferring the mappings between words and their referents is a difficult problem that all language learners face. Similarly, learning which word orders are permitted in one's language is one of the first grammatical learning tasks these same learners must solve. We present a modeling framework which addresses simple versions of both of these problems by using the joint information in each to bootstrap the other. We discover that these two distinct learning tasks may be easier to solve jointly because of the way in which the inferences in one problem constrain the inferences in the other.

Keywords: word learning; word order; Bayesian models; mutual constraint; reference; linguistics

Introduction

The language-learning child is faced with two simultaneous acquisition problems: acquiring the (semantic) rules that map the words she hears onto the objects and actions she perceives, and acquiring the (syntactic) rules that govern how those words should be combined to make grammatical sentences. Both are difficult learning problems in their own right, and have been the topic of considerable research.

Determining the meaning of words on the basis of real-life observational evidence is quite difficult (Gillette, Gleitman, Gleitman, & Lederer, 1991), in part because of the inherent ambiguity of words, in part because the number of potential meanings is logically underconstrained (Quine, 1960). While it may be that the identification of a word's referent is made easier by pre-existing biases (Markman, 1990), recent research has also suggested several methods by which children could explicitly learn which objects or actions a particular word refers to. For instance, social cues such as pointing or gaze (Frank, Goodman, & Tenenbaum, 2007) can assist the learner, as can a sensitivity to the statistics of cross-situational word learning (Frank et al., 2007; Yu & Smith, 2008) and the ability to form theories about the abstract rules that govern the mapping of words onto categories (e.g., Kemp, Perfors, & Tenenbaum, 2007). Experiments and computational modeling suggest that the difficulties and ambiguities inherent in cross-situational word learning can be at least partially alleviated by these techniques.

Acquiring the rules of syntax is also a famously difficult problem. Even if we restrict ourselves to more tractable subproblems – for instance, the acquisition of word order – the empirical data present some difficult issues. Children make few mistakes in word order when they start combining words (e.g., Brown, 1973), and even children who do not combine words have been shown to make use of word order knowledge in a comprehension task (Hirsh-Pasek & Golinkoff, 1996). This suggests that word order knowledge is acquired very early, but it is not clear how it is done. Although learning word order is a more limited problem than learning the referents of words, since there are fewer possible solutions to the problem, it is still a puzzle how it can be done so quickly. Some have suggested that prosodic bootstrapping may explain a related problem, the acquisition of head direction (Christophe, Nespor, Guasti, & Ooyen, 2003). Although this requires the assumption of innate knowledge of the mapping principles between prosodic cues and head direction, and does not address the acquisition of word order itself, prosodic bootstrapping may play an important role.

In this paper we propose that both of these acquisition problems can be made more tractable by addressing them jointly. On the one hand, if the learner believes that word orderings tend to be consistent, constraints are imposed on the manner in which words may be mapped onto entities in the world. On the other hand, even knowing a few word meanings is enough to provide a great deal of evidence about word order. These intuitions suggest that viewing the problem as a *joint* acquisition problem can make both individual problems easier. While in one sense the proposition may appear counterintuitive – after all, there is in some sense 'more' to learn in the joint problem – to the extent that each problem mutually constrains the other, the acquisition problem should be made *less* difficult, rather than more. This basic idea is not a new one: for instance, earlier work noted its potential (Siskind, 1990, 1991). However, performing inferences about both syntactic and semantic information was beyond the computational capabilities of the time, and in practice, that work simply demonstrated that hardwired syntactic information could make the learning of semantics easier. Our research goes beyond this work in two ways: first, because we demonstrate that truly *joint* inference, in which both aspects of the problem mutually bootstrap each other, can make the learning problem easier; and second, because the syntactic information is simpler and sparser (word order rather than X-bar theory or richer grammatical knowledge). Our study presents two models that seek to establish word-referent mappings on the basis of cross-situational learning statistics: one model also seeks to acquire word order, and uses this to assist word-referent mapping learning, and one does not. We demonstrate that solving the joint acquisition problem results in more rapid learning of word reference.

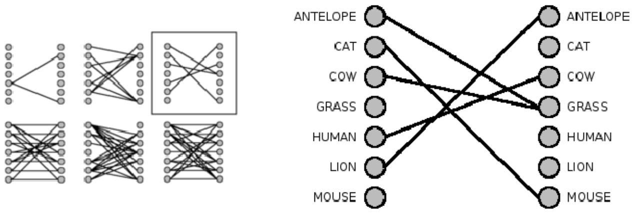

Figure 1: A simple example world, consisting of 7 objects corresponding to common animals and 6 relations. The leftmost portion of the figure shows representations of the relations. The top left relation, which corresponds to the concept of EATS, is enlarged in the rightmost portion of the figure for clarity. The object labels in this portion are for the reader's convenience: they are not inherent properties of the objects and are not visible to our model.

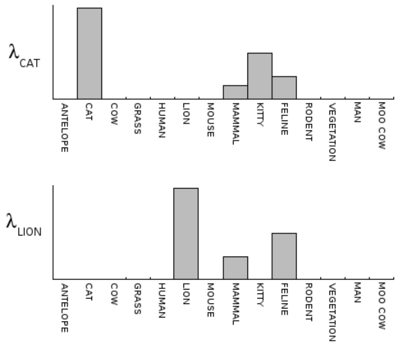

Figure 2: Naming distributions for two objects in our toy world, demonstrating synonymy (both objects have more than one word with non-zero naming probability) and polysemy (the words "feline" and "mammal" have non-zero naming probability for both objects).

A Simple Language & World Setup

Our models consider a learner who exists in a physical world of objects and inter-object relations. The learner is attempting to acquire a language (consisting of word order knowledge and a lexicon of word-world mappings) through exposure to concurrent observations of the world and linguistic input. Though heavily simplified, it is intended as a first-order approximation to the acquisition problem facing children.

The World

Formally, our world is specified in terms of a set of m objects, $\mathcal{O} = \{o_1, \dots, o_m\}$, and a set of n relations that exist between those objects, $\mathcal{R} = \{r_1, \dots, r_n\}$. Each relation is a function defined for pairs of objects (i.e., $r_i \subseteq \mathcal{O} \times \mathcal{O}$); if the relation holds for two objects $r(o_1, o_2)$ is true. Not all true things are equally likely to be observed: if BITES(o_1, o_2) indicates that object 1 is able to bite object 2, then the specific observation BITES(*dog*, *man*) will be made much more frequently than BITES(*man*, *dog*). We formalize this notion by equipping the world with a probability distribution $\Phi(\cdot)$ over observed relationships. The learner's physical observations are generated from $\Phi(\cdot)$, and consist only of true statements about the world, but some things are seen much more often than others. Figure 1 shows a diagrammatic representation of a simple example world involving 7 objects and 6 relations. The structure of one of the relations, corresponding to the concept of EATS($\cdot$,$\cdot$) is magnified.

The Language

The language component of the modeling involves a probabilistic lexicon with a vocabulary of v words, $\mathcal{V} = \{w_1, \dots, w_v\}$. For every object or relation x in the world, there is a *naming distribution* λ_x over the vocabulary (i.e., $\lambda_x : \mathcal{V} \to [0,1]$). We denote the set of all naming distributions by Λ. A naming distribution is essentially the map between items in the world and the words for those items; it assigns higher probability to those words more likely to be used as names for the relevant object. For instance, if the object x corresponds to the entity *cat*, the distribution λ_x should assign the most probability to the word "cat", a substantial amount to words such as "kitty" or "pet", a small but non-zero amount to "feline" and no probability to the words "monkey", "peanut" or "indigo". To make this scheme more explicit, Figure 2 shows two examples of a naming distribution; it illustrates how this scheme permits both synonymy and polysemy, thus reproducing some of the factors which complicate word learning in the real world.

In addition to the probabilistic lexicon, our simplified language model includes the concept of word order. We consider a set of six word orders corresponding to the six possible ways of ordering subjects, verbs and objects.[1] The word order in our language is specified by a probability distribution Θ over the set of these six possible word orders. As an illustration, we might think of the English language as assigning 80% probability to the SVO ordering, 20% to the OVS order, and 0% to all other word orders. This distribution encodes a strong preference for active voice, allows the occasional use of passive voice, and indicates that the other four word orderings are ungrammatical.[2]

The Nature of the Input

In our simulations we generate a collection of observations from the world and corresponding data from the language, and the learner's task is to use this input to infer the correct underlying naming distribution λ_x for each object and relation – and perhaps, jointly, to infer the correct word order Θ for the language. Formally, the input available to the learner, $\mathcal{D}$, consists of observations of relations and objects, $z = r(o_1, o_2)$, which are drawn from $\Phi(\cdot)$, each of which is paired with a three-word linguistic utterance, $w = w_1 w_2 w_3$, which is generated by randomly selecting a word for each of r, o_1 and o_2 from the appropriate naming distributions and combining them to form w using a word order θ drawn from the language's word order distribution Θ. For instance, if the selected word order is $\theta =$ SVO then $w_1 \sim \lambda_{o_1}$, $w_2 \sim \lambda_r$ and $w_3 \sim \lambda_{o_2}$. Each data point in $\mathcal{D}$ corresponds to a coupled observation-sentence pairing generated in this way, i.e. $\mathcal{D} = \{d_1 = (z_1, w_1), d_2 = (z_2, w_2), \dots\}$. Each d_i implicitly has a word order variable θ_i associated with it, which is not ob-

[1] This includes SVO, SOV, VSO, VOS, OSV, and OVS.

[2] Note that we do not, in fact, encode a preference for any *particular* word order – whether found in English or not – into the model. We merely allow the model to postulate that some orders will turn out to be more common than others in the target language.

Table 1: Example input data $\mathcal{D}$. Each row represents a single datum, coupling a relational observation z with a linguistic one w.

Relational observation (z)	Linguistic utterance (w)
EAT(*cat*, *mouse*)	“cat eat rodent”
CHASE(*lion*, *antelope*)	“lion chase prey”
EAT(*cow*, *grass*)	“cow consume grass”
EAT(*antelope*, *grass*)	“antelope eat grass”

servable by the learner.

Note that our linguistic input differs from ‘real’ input in that we give no regard to functional words such as “a”, “the” or “this”. Filtering complete sentences in this way seems reasonable given that young infants are capable of making the distinction between function and content words on the basis of frequency and prosody (e.g., Jusczyk, 1997). We are also assuming that a language learner is able to unambiguously associate each linguistic utterance with a relational observation, which may rely on the use of cues like gaze. (We discuss this oversimplification later in the paper).

A brief example of the data is given in Table 1. Note that the learner does not have direct knowledge of the relational structure shown in Figure 1, the correct naming distributions in Figure 2, or knowledge of which elements in the observation map onto which words in the linguistic utterance: everything must be inferred from the data in $\mathcal{D}$.

Methodology

Models

The main motivation for our research was to explore to what extent two difficult acquisition problems – establishing word reference, and learning word order – could each be made easier by attempting to solve them jointly. To that end, we compare two word learning models that differ in their ability to acquire word order information. Both models seek to infer the correct naming distributions Λ and are presented with data $\mathcal{D}$. Each individual datapoint d consists of coupled observations and three-word utterances (z, w). The difference between the models is that the baseline model, which we call M_B, assumes that there is no consistent word order in the language; the word-order learning model, which we call M_{WO}, assumes that the language has a consistent distribution over word orders, Θ, and seeks to learn that as well as the naming distribution. While it may seem cognitively implausible that real language learners maintain some mental representation of a complete probability distribution over possible labels for each object or concept they encounter, as both our models do, this idea has received some empirical credibility from recent experimental work (Vouloumanos, 2008); additionally, it may not be necessary to have a precisely accurate probability distribution in order to receive substantial benefit from joint learning (although that is a topic for further research).

To elaborate on the difference between models, the baseline model M_B implicitly assumes that the distribution Θ over word orderings is perfectly uniform. That is, given the coupled $z = r(o_1, o_2)$ and $w = w_1 w_2 w_3$, it does assume that each word refers to precisely one of the three relations or objects, but does not try to learn any consistent mappings – *a priori*, w_1 is just as likely to refer to the relation r as it is to one of the objects o_1 or o_2, and the same is true of w_2 and w_3. This forces the model to rely only upon the concurrence of relations or objects and words in attempting to estimate the set of naming distributions Λ.

The estimates of the naming distributions, which we denote by $\hat{\Lambda}$, are calculated via Bayesian inference over the space of possible naming distributions; a symmetric Dirichlet distribution with parameter α serves as our prior for each of the $\hat{\lambda}_x$. We perform the inference numerically using Gibbs sampling, a common and convenient form of Markov Chain Monte Carlo[3] (Gilks, Richardson, & Spiegelhalter, 1996). This involves iteratively assigning a word order variable θ_i to each data point d_i in $\mathcal{D}$. Each of these assignments is made randomly using a probability distribution conditioned on all the other assignments: this full conditional distribution and other technical details are available in the appendix. For now it will suffice to say that the probability of assigning a particular word order θ to a given data point is proportional only to its consistency with other assignments; in other words, the model prefers words to have few meanings and meanings to be associated with few words. Note that although the model does learn word order assignments θ_i, it does not learn any general rules about word order that hold *across* utterances. The θ_i values that it learns correspond only to the mapping from the particular words in the utterance w_i to the entities in the observation z_i.

The word-order learning model M_{WO} is identical to the baseline model except that is assumes that word order tends to be consistent across all utterances. The learner thus aims to estimate some explicit, non-uniform word order distribution $\hat{\Theta}$. Once again, we model this using Bayesian inference, assuming that the learner places a symmetric Dirichlet prior distribution with parameter β over the possible word order distributions. In this model, the probability of assigning a particular word order θ to a given data point is dependent on the consistency of its word-world assignments (as in M_B), as well as the consistency of word orderings across data points. Technical details for both models, including the full conditional distribution, are available in the appendix.

Data sets

Simulated data sets are created based on the generative process detailed earlier. To explore how performance changes as a function of the quantity of data, we create a series of data sets $\mathcal{D}$ with varying numbers of observation-sentence pairs. Data sets with more data points are generated by adding additional points to the smaller data sets. All results are averaged over 10 different data sets at each size; each set was generated using different random values of Φ and Λ.

Results

The task of our learner was to make reasonable inferences about the likely referents of each of the words in the language, as well as, in the case of M_{WO}, to determine the probable word order in the language. Figure 3 depicts the rate of acquisition

[3] We do *not* suggest that child language learners literally implement Gibbs sampling or Bayesian inference. We use these tools as models of “ideal learning” in order to explore whether mutual constraint in this task is sufficient to make the joint learning problem substantially easier, and what could be learned in principle.

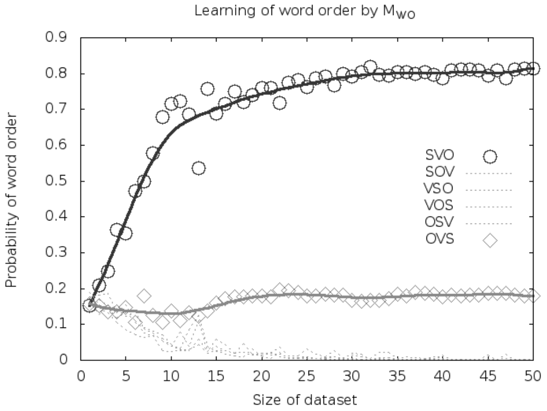

Figure 3: Inferred word order probabilities by model M_W for various sized data sets. The world has 20 objects, 10 relations and 50 words.

of word order by M_{WO} as the quantity of data increases. The correct word order distribution Θ for this data assigned probability 0.8 to the word order SVO and 0.2 to OVS, with all other orderings receiving zero probability.[4] It is evident that only a small amount of data is necessary before the model accurately infers the correct word order – Figure 3 shows that the inferred probabilities are essentially perfect with a data set size of 30 or above, and are approximately correct with as small a data set size as 15. In a sense this is not surprising, given that there are only six possibilities to choose from, but it is noteworthy in light of children's early acquisition of word order. We note that for the simulations which produced this data, we used a Dirichlet distribution parameter of $\beta = 1$ for the prior estimate of Θ. Such a value provides no bias in the direction of sparsity or non-sparsity. The fact that word order can be acquired quickly from so few 'coupled' data despite the lack of bias may suggest no need to hypothesize that children are born with strong innate constraints on word ordering to explain their rapid acquisition.[5]

How well does the model acquire the correct word-world mappings? We assess this by calculating the accuracy of the inferred naming distributions for each object in the world. Because the learner induces entire naming *distributions* λ_x for each object x, rather than mappings to a single lexical item, calculating this is not completely straightforward. We measured accuracy in two ways:

1. By calculating the average Kullback-Leibler divergence[6] between actual naming distributions and their corresponding inferred naming distributions.

2. By calculating the proportion of learned naming distributions that have the correct modal mapping: a distribution that predicted that CAT mapped onto "cat" 60% of the time and "fiberglass" 40% of the time would count as correct, and one that predicted the reverse would not.

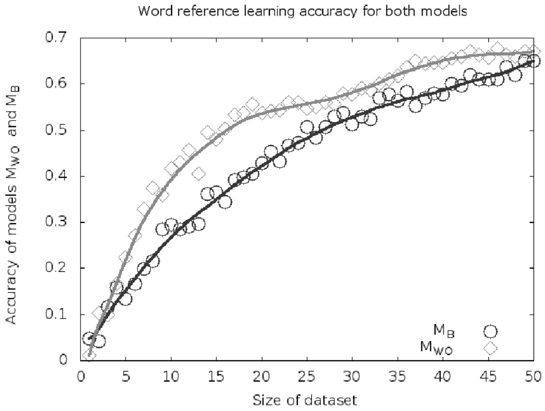

Figure 4: Accuracy of models M_B and M_{WO}, approximated by proportion of inferred naming distributions with correct means, based on data set size. The world has 20 objects, 10 relations and 50 words. M_{WO} is shown with gray diamonds and M_B with black circles.

These two measures were chosen to address the conflicting criteria of intuitive interpretability (which is satisfied by calculating the proportion of learned distributions with correct modes) and accuracy (which is better satisfied by calculating KL divergence). Since we found no qualitative difference between the results depending on which measure was used, we present all results here in the second, more intuitive format.

Figure 4 shows the accuracy of M_B and M_{WO}, as measured by the proportion of naming distributions with correct modes, for data set sizes ranging from 1 to 50. These datasets were generated using a simple world consisting of 20 objects, 10 relations and 50 words. For both models, accuracy increases as the quantity of data increases, and accuracy is overall quite high: after observing only 20 utterance-observation pairs, the word-order learning model M_{WO} has found the correct referent for over 50% (i.e., over 15 of the 30) of the relations and objects. Even the baseline model M_B has acquired around 40%, which provides further evidence for the observation, suggested by other researchers, that learning of reference can be greatly facilitated by the use of cross-situational statistical information (Frank et al., 2007; Yu & Smith, 2008).

More interestingly, we also observe that M_{WO} outperforms the baseline M_B; this is shown more clearly in Figure 5, which shows the difference in accuracy between the two models. It is clear that jointly learning word order offers a significant advantage, especially when the amount of data is small. This advantage decreases as the data set increases in size, which is to be expected: in the limit, the high quantity of correlated cross-situational information should suffice to overcome any ambiguities in reference. Importantly, smaller data sets are of special interest to us, since they more closely approximate the inference problem facing the child, who receives quite sparse data relative to the amount to be learned in the world, and shows rapid learning in that situation. Our result suggests that children may be able to use inferences about word order – which are supported quite early – to bootstrap their inferences about word reference.

To what extent are these results due to the fact that our toy

[4] Each of our simulations were performed with two correct word order distributions, one which placed all probability on a single word order and one which split the probability between two orderings with probabilities 0.8 and 0.2. No qualitative differences in our results were observed. All figures presented in this paper correspond to data generated with the bimodal distribution.

[5] We also tested the $\beta = 0.01$ case, which encodes a strong bias toward sparsity. This made little qualitative difference to the results.

[6] The KL divergence between two distributions P and Q defined on the set X is given by $D_{KL}(P||Q) = \sum_{x \in X} P(x) \ln(P(x)/Q(x))$.

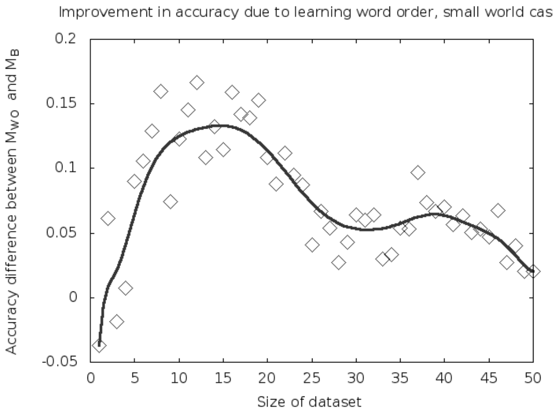

Figure 5: Accuracy benefit to joint learning in a small world. Comparison of the baseline model (M_B) with the word-order learning model (M_{WO}) in terms of accuracy of acquiring the correct word-world mappings in a small world with 20 objects, 10 relations, and 50 words. The *y* axis shows the increase in accuracy that comes from jointly learning word order as well as reference alone. Model M_{WO} clearly outperforms M_B, particularly when there is little data.

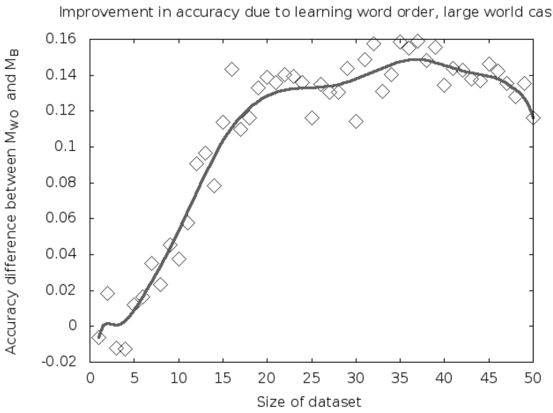

Figure 6: Accuracy benefit to joint learning in a large world. Comparison of (M_B) with (M_{WO}) in terms of accuracy in a world with 80 objects, 40 relations, and 200 words. Once again the joint model M_{WO} clearly outperforms the baseline M_B. In the larger world the duration of the effect appears to be greater.

world is relatively small, with few objects, words, and relations? While constructing a world of the same complexity that the child faces is beyond our purview, we address the issue of scalability by presenting the same models with data from a substantially larger world (80 objects, 40 relations, and 200 words). Figure 6 depicts the same accuracy advantage of M_{WO} over M_B as for smaller amounts of data, but that advantage is retained for longer. This is sensible because in a larger world, significantly more data is required before the information conveyed by cross-situational correlation information alone is sufficiently saturated to negate the advantage of also being able to use word order. This suggests that the extreme simplicity of our small world compared to the real world has not exaggerated the strength of the advantage of joint learning; in fact, it may have underestimated it. In a world as large and complicated as the real world, being able to rely on inferences about word order to figure out the meaning of the words in the sentence may be of significant benefit.

Discussion

This work demonstrates that two distinct language acquisition problems – learning word reference and inducing word order – can be made easier by addressing them jointly. While in some sense this is counter-intuitive, since in the joint problem there is 'more' to be learned, we suggest that the joint problem is in fact easier because each problem constrains the other. Knowing that verbs tend to be first can enable a learner to map the word "glim" in the sentence "glim torg nim" onto the action in the world; conversely, knowing that "glim" refers to a kind of biting action can enable a learner to infer, upon hearing the same sentence, that words denoting actions may come first. This is sensible, but has not until now been supported by quantitative analysis.

Our world and the learning situation are in many ways vastly oversimplified versions of the task facing the child learning language. Our goal here is not to argue that children approach the situation in precisely the way our models do, but rather to lend some empirical support to the notion that joint learning of two complicated tasks can make both tasks easier. We suggest that many types of inference – which classic learnability analysis would suggest are too difficult for children to acquire as rapidly as they do – may be significantly easier when conceptualized as a joint problem in language and higher-order cognition. Moreover, by constructing models that explicitly handle the joint inference problem as well as models for each of the individual ones, we can begin to quantify both the qualitative and quantitative features of the speedup effect.

More broadly, this modeling framework can be expanded in interesting ways to explore problems of more complexity and, thus, greater applicability to the situations faced by child learners. The model currently assumes that all data consists of joint utterances and observations of the world – yet often children are in situations where they observe objects and events happening but receive no linguistic input, or where they hear sentences that have no apparent connection to the events in the world. What happens if the model is presented with data sets consisting of all three kinds of data? Preliminary indications suggest that the advantage of joint learning still exists – indeed, the learner is still able to leverage some information out of the singleton data: for instance, observations of events without language still provide evidence about what kinds of events are more or less likely. Future work will explore this issue in more detail.

Another shortcoming of the current modeling framework is that it makes certain implicit assumptions about the nature of the knowledge the learner starts with. Our word-order learning component assumes that the learner already has concepts for subjects, objects and verbs, and that languages may differ in how those are ordered. While there is some evidence that notions of agency and objecthood form a core part of cognition from infancy (e.g., Spelke & Kinzler, 2007), an interesting extension to this analysis would be to present input consisting of items with features, and explore whether the model could induce the notions of subject, object and verb, based on a presumption that word order is consistent and that words map onto things in the world. This framework is also easily extendible to address the acquisition of more complicated syntactic knowledge: for instance, the realization that in some languages it is permissible to optionally drop subject pronouns. In other languages, word order plays a much less important role than it does in English: this information is conveyed by other means, such as morphological inflection. This, too, could be added to our model, in addition to a word-order learning component. One would expect that an effective learner would learn to make use of whichever kind of information was most informative, although further work is necessary to explore whether this expectation is correct, and how much different types of information help with the overall learning problem.

In general, the analysis here provides a framework for investigating how the joint acquisition of distinct pieces of knowledge can make the acquisition of each individual piece easier. Our results suggest that classic learnability problems, which often presume that information is acquired in isolation, may not always apply to the situation facing the child learner.

Acknowledgements

DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). We thank three anonymous reviewers for their comments.

References

Brown, R. (1973). *A first language*. Cambridge, MA: Havard University Press.

Christophe, A., Nespor, M., Guasti, M., & Ooyen, B. (2003). Prosodic structure and syntactic acquisition: The case of the head-direction parameter. *Developmental Science, 6*, 211–220.

Frank, M., Goodman, N., & Tenenbaum, J. (2007). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems, 20*.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1991). Human simulations of vocabulary learning. *Cognition*, *73*, 153–176.

Hirsh-Pasek, K., & Golinkoff, R. (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.

Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Siskind, J. (1990). Acquiring core meanings of words, represented as Jackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. *ACL*.

Siskind, J. (1991). Dispelling myths about language bootstrapping. *AAAI Workshop on Machine Learning of Natural Language and Ontology*, 157–64.

Spelke, E., & Kinzler, K. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–749.

Yu, C., & Smith, L. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568.

Appendix

For all Gibbs samplers used in our models, we employ an initial ‘burn in’ period of 1000 iterations and then generate our estimate histograms using 500 samples, with an inter-sample lag of 100 iterations.

Model M_B

The full conditional distribution for the word order θ_i (assigned to the ith component of $\mathcal{D}$, d_i), is given below, where we denote the relational component of d_i by $z_i = r(s,o)$, the linguistic component by $w_i = w_1 w_2 w_3$, and by $\boldsymbol{\theta}_{-i}$ the set of all *other* word order assignments:

$$\begin{aligned} P(\theta_i \,|\, \boldsymbol{\theta}_{-i}, \mathcal{D}\,) &\propto P(\theta_i \,|\, \boldsymbol{\theta}_{-i}) P(w_i | z_i, \theta_i) \\ &\propto \hat{\lambda}_r(\textsc{rel}_{\theta_i}(w_i))\hat{\lambda}_r(\textsc{sub}_{\theta_i}(w_i))\hat{\lambda}(\textsc{obj}_{\theta_i}(w_i)) \end{aligned}$$

Note that the term $P(\theta_i \,|\, \boldsymbol{\theta}_{-i})$ has been absorbed by the proportionality, by virtue of the assumption that it is a constant (i.e., 1/6). The functions $\textsc{rel}_\theta, \textsc{sub}_\theta$ and $\textsc{obj}_\theta$ are defined for each possible value of θ so that, given the input $w = w_1 w_2 w_3$, they return the word which corresponds to the relation, subject and object, respectively, given the particular word order θ. For instance, if $\theta = \text{SVO}$, then $\textsc{rel}_\theta(w) = w_2, \textsc{sub}_\theta(w) = w_1$ and $\textsc{obj}_\theta(w) = w_3$.

Here $\hat{\lambda}_x$, for $x = s, r, o$ are our inferred approximations to the relevant naming distributions. At any iteration, these approximations are given by the following expression, which is arrived at by applying Bayes’ law and the use of the same symmetric Dirichlet distribution prior, with parameter α, for all naming distributions:

$$\hat{\lambda}_x(w) = \frac{n_{xw}(x,w) + \alpha}{n_x(x) + v\alpha}$$

Here $n_{xw}(x,w)$ counts the number of data points in which the relational component contains the relation or object x, the linguistic component contains the word w, and the word order assigned to the utterance is such that w is understood to be a name for x. The term $n_x(x)$ counts the number of observations z which involve x. We have used $\alpha = 0.01$ in our simulations, which represents a strong prior bias toward sparsity of the naming distributions.

Model M_{WO}

Reusing our notation from model M_B, the full conditional distribution for word order assignments in M_{WO} is:

$$\begin{aligned} P(\theta_i \,|\, \boldsymbol{\theta}_{-i}, \mathcal{D}\,) &\propto P(\theta \,|\, \boldsymbol{\theta}_{-i}) P(w_i \,|\, z_i, \theta_i) \\ &\propto \hat{\Theta}(\theta_i) P(w_i \,|\, z_i, \theta_i) \end{aligned}$$

Here the rightmost term, representing the likelihood of w_i being generated as a description of z_i given the word order θ_i, is exactly as before. The leftmost term $\hat{\Theta}$ is our inferred approximation to the word order distribution, which at any iteration is given by:

$$\hat{\Theta}(\theta) = \frac{n_\theta(\theta) + \beta}{|\mathcal{D}| + 6\beta}$$

Here $n_\theta(\theta)$ counts the number of linguistic components of $\mathcal{D}$ which have been assigned the word order θ.