Exploiting social information in grounded language learning via grammatical reductions

Mark Johnson
Department of Computing
Macquarie University
Sydney, Australia
Mark.Johnson@MQ.edu.au

Katherine Demuth
Department of Linguistics
Macquarie University
Sydney, Australia
Katherine.Demuth@MQ.edu.au

Michael Frank
Department of Psychology
Stanford University
Stanford, California
mcfrank@Stanford.edu

Abstract

This paper uses an unsupervised model of grounded language acquisition to study the role that social cues play in language acquisition. The input to the model consists of (orthographically transcribed) child-directed utterances accompanied by the set of objects present in the non-linguistic context. Each object is annotated by *social cues*, indicating e.g., whether the caregiver is looking at or touching the object. We show how to model the task of inferring which objects are being talked about (and which words refer to which objects) as standard grammatical inference, and describe PCFG-based *unigram* models and adaptor grammar-based *collocation* models for the task. Exploiting social cues improves the performance of all models. Our models learn the relative importance of each social cue jointly with word-object mappings and collocation structure, consistent with the idea that children could discover the importance of particular social information sources during word learning.

1 Introduction

From learning sounds to learning the meanings of words, social interactions are extremely important for children's early language acquisition (Baldwin, 1993; Kuhl et al., 2003). For example, children who engage in more joint attention (e.g. looking at particular objects together) with caregivers tend to learn words faster (Carpenter et al., 1998). Yet computational or formal models of social interaction are rare, and those that exist have rarely gone beyond the stage of cue-weighting models. In order to study the role that *social cues* play in language acquisition, this paper presents a structured statistical model of grounded learning that learns a mapping between words and objects from a corpus of child-directed utterances in a completely unsupervised fashion. It exploits five different *social cues*, which indicate which object (if any) the child is looking at, which object the child is touching, etc. Our models learn the salience of each social cue in establishing reference, relative to their co-occurrence with objects that are not being referred to. Thus, this work is consistent with a view of language acquisition in which children *learn to learn*, discovering organizing principles for how language is organized and used socially (Baldwin, 1993; Hollich et al., 2000; Smith et al., 2002).

We reduce the grounded learning task to a grammatical inference problem (Johnson et al., 2010; Börschinger et al., 2011). The strings presented to our grammatical learner contain a prefix which encodes the objects and their social cues for each utterance, and the rules of the grammar encode relationships between these objects and specific words. These rules permit every object to map to every word (including function words; i.e., there is no "stop word" list), and the learning process decides which of these rules will have a non-trivial probability (these encode the object-word mappings the system has learned).

This reduction of grounded learning to grammatical inference allows us to use standard grammatical inference procedures to learn our models. Here we use the *adaptor grammar* package described in Johnson et al. (2007) and Johnson and Goldwater (2009) with "out of the box" default settings; no parameter tuning whatsoever was done. Adaptor grammars are a framework for specifying hierarchical non-parametric models that has been previously used to model language acquisition (Johnson, 2008).

Social cue	Value
child.eyes	objects child is looking at
child.hands	objects child is touching
mom.eyes	objects care-giver is looking at
mom.hands	objects care-giver is touching
mom.point	objects care-giver is pointing to

Figure 1: The 5 social cues in the Frank et al. (to appear) corpus. The value of a social cue for an utterance is a subset of the available topics (i.e., the objects in the non-linguistic context) of that utterance.

A semanticist might argue that our view of referential mapping is flawed: full noun phrases (e.g., *the dog*), rather than nouns, refer to specific objects, and nouns denote properties (e.g., *dog* denotes the property of being a dog). Learning that a noun, e.g., *dog*, is part of a phrase used to refer to a specific dog (say, Fido) does not suffice to determine the noun's meaning: the noun could denote a specific breed of dog, or animals in general. But learning word-object relationships is a plausible first step for any learner: it is often only the contrast between learned relationships and novel relationships that allows children to induce super- or sub-ordinate mappings (Clark, 1987). Nevertheless, in deference to such objections, we call the object that a phrase containing a given noun refers to the *topic* of that noun. (This is also appropriate, given that our models are specialisations of topic models).

Our models are intended as an "ideal learner" approach to early social language learning, attempting to weight the importance of social and structural factors in the acquisition of word-object correspondences. From this perspective, the primary goal is to investigate the relationships between acquisition tasks (Johnson, 2008; Johnson et al., 2010), looking for synergies (areas of acquisition where attempting two learning tasks jointly can provide gains in both) as well as areas where information overlaps.

1.1 A training corpus for social cues

Our work here uses a corpus of child-directed speech annotated with social cues, described in Frank et al. (to appear). The corpus consists of 4,763 orthographically-transcribed utterances of caregivers to their pre-linguistic children (ages 6, 12, and 18 months) during home visits where children played with a consistent set of toys. The sessions were video-taped, and each utterance was annotated with the five social cues described in Figure 1.

Each utterance in the corpus contains the following information:

- the sequence of orthographic words uttered by the care-giver,
- a set of *available topics* (i.e., objects in the non-linguistic objects),
- the values of the social cues, and
- a set of *intended topics*, which the care-giver refers to.

Figure 2 presents this information for an example utterance. All of these but the intended topics are provided to our learning algorithms; the intended topics are used to evaluate the output produced by our learners.

Generally the intended topics consist of zero or one elements from the available topics, but not always: it is possible for the caregiver to refer to two objects in a single utterance, or to refer to an object not in the current non-linguistic context (e.g., to a toy that has been put away). There is a considerable amount of anaphora in this corpus, which our models currently ignore.

Frank et al. (to appear) give extensive details on the corpus, including inter-annotator reliability information for all annotations, and provide detailed statistical analyses of the relationships between the various social cues, the available topics and the intended topics. That paper also gives instructions on obtaining the corpus.

1.2 Previous work

There is a growing body of work on the role of social cues in language acquisition. The language acquisition research community has long recognized the importance of social cues for child language acquisition (Baldwin, 1991; Carpenter et al., 1998; Kuhl et al., 2003).

Siskind (1996) describes one of the first examples of a model that learns the relationship between words and topics, albeit in a non-statistical framework. Yu and Ballard (2007) describe an associative learner that associates words with topics and that exploits prosodic as well as social cues. The relative importance of the various social cues are specified a priori in their model (rather than learned, as they are here), and unfortunately their training corpus is not available. Frank et al. (2008) describes a Bayesian model that learns the relationship between words and topics, but the version of their model that included social cues presented a number of challenges for inference. The unigram model we describe below corresponds most closely to the Frank

.dog # .pig child.eyes mom.eyes mom.hands # ## wheres the piggie

Figure 2: The photograph indicates non-linguistic context containing a (toy) pig and dog for the utterance *Where's the piggie?*. Below that, we show the representation of this utterance that serves as the input to our models. The prefix (the portion of the string before the "##") lists the available topics (i.e., the objects in the non-linguistic context) and their associated social cues (the cues for the pig are *child.eyes*, *mom.eyes* and *mom.hands*, while the dog is not associated with any social cues). The intended topic is the pig. The learner's goals are to identify the utterance's intended topic, and which words in the utterance are associated with which topic.

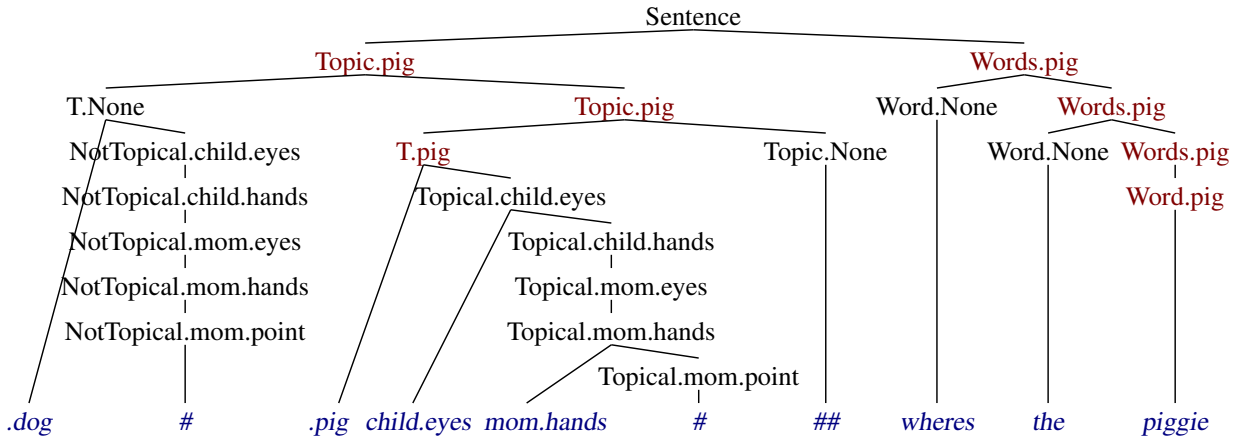

Figure 3: Sample parse generated by the Unigram PCFG. Nodes coloured red show how the "pig" topic is propagated from the prefix (before the "##" separator) into the utterance. The social cues associated with each object are generated either from a "Topical" or a "NotTopical" nonterminal, depending on whether the corresponding object is topical or not.

et al. model. Johnson et al. (2010) reduces grounded learning to grammatical inference for adaptor grammars and shows how it can be used to perform word segmentation as well as learning word-topic relationships, but their model does not take social cues into account.

2 Reducing grounded learning with social cues to grammatical inference

This section explains how we reduce ground learning problems with social cues to grammatical inference problems, which lets us apply a wide variety of grammatical inference algorithms to grounded learning problems. An advantage of reducing grounded learning to grammatical inference is that it suggests new ways to generalise grounded learning models; we explore three such generalisations here. The main challenge in this reduction is finding a way of expressing the non-linguistic information as part of the strings that serve as the grammatical inference procedure's input. Here we encode the non-linguistic information in a "prefix" to each utterance as shown in Figure 2, and devise a grammar such that inference for the grammar corresponds to learning the word-topic relationships and the salience of the social cues for grounded learning.

All our models associate each utterance with zero or one topics (this means we cannot correctly analyse utterances with more than one intended topic). We analyse an utterance associated with zero topics as having the special topic None, so we can assume that every utterance has exactly one topic. All our grammars generate strings of the form shown in Figure 2, and they do so by parsing the prefix and the words of the utterance separately; the top-level rules of the grammar force the same topic to be associated with both the prefix and the words of the utterance (see Figure 3).

2.1 Topic models and the unigram PCFG

As Johnson et al. (2010) observe, this kind of grounded learning can be viewed as a specialised kind of topic inference in a topic model, where the utterance topic is constrained by the available objects (possible topics). We exploit this observation here using a reduction based on the reduction of LDA topic models to PCFGs proposed by Johnson (2010). This leads to our first model, the unigram grammar, which is a PCFG.[1]

$\mathrm{Sentence} \rightarrow \mathrm{Topic}_t\ \mathrm{Words}_t$	$\forall t \in T'$
$\mathrm{Topic}_{\mathrm{None}} \rightarrow \#\#$	
$\mathrm{Topic}_t \rightarrow \mathrm{T}_t\ \mathrm{Topic}_{\mathrm{None}}$	$\forall t \in T'$
$\mathrm{Topic}_t \rightarrow \mathrm{T}_{\mathrm{None}}\ \mathrm{Topic}_t$	$\forall t \in T$
$\mathrm{T}_t \rightarrow t\ \mathrm{Topical}_{c_1}$	$\forall t \in T$
$\mathrm{Topical}_{c_i} \rightarrow (c_i)\ \mathrm{Topical}_{c_{i+1}}$	$i = 1, \ldots, \ell - 1$
$\mathrm{Topical}_{c_\ell} \rightarrow (c_\ell)\ \#$	
$\mathrm{T}_{\mathrm{None}} \rightarrow t\ \mathrm{NotTopical}_{c_1}$	$\forall t \in T$
$\mathrm{NotTopical}_{c_i} \rightarrow (c_i)\ \mathrm{NotTopical}_{c_{i+1}}$	$i = 1, \ldots, \ell - 1$
$\mathrm{NotTopical}_{c_\ell} \rightarrow (c_\ell)\ \#$	
$\mathrm{Words}_t \rightarrow \mathrm{Word}_{\mathrm{None}}\ (\mathrm{Words}_t)$	$\forall t \in T'$
$\mathrm{Words}_t \rightarrow \mathrm{Word}_t\ (\mathrm{Words}_t)$	$\forall t \in T$
$\mathrm{Word}_t \rightarrow w$	$\forall t \in T', w \in W$

Figure 4: The rule schema that generate the unigram PCFG. Here $(c_1, \ldots, c_\ell)$ is an ordered list of the social cues, T is the set of all non-None available topics, $T' = T \cup \{\mathrm{None}\}$, and W is the set of words appearing in the utterances. Parentheses indicate optionality.

Figure 4 presents the rules of the unigram grammar. This grammar has two major parts. The rules expanding the Topic_t nonterminals ensure that the social cues for the available topic t are parsed under the $\mathrm{Topical}$ nonterminals. All other available topics are parsed under $\mathrm{T}_{\mathrm{None}}$ nonterminals, so their social cues are parsed under $\mathrm{NotTopical}$ nonterminals. The rules expanding these non-terminals are specifically designed so that the generation of the social cues corresponds to a series of binary decisions about each social cue. For example, the probability of the rule

$$\mathrm{Topical}_{child.eyes} \rightarrow .\textit{child.eyes}\ \mathrm{Topical}_{child.hands}$$

is the probability of an object that is an utterance topic occuring with the *child.eyes* social cue. By estimating the probabilities of these rules, the model effectively learns the probability of each social cue being associated with a $\mathrm{Topical}$ or a $\mathrm{NotTopical}$ available topic, respectively.

The nonterminals Words_t expand to a sequence of Word_t and $\mathrm{Word}_{\mathrm{None}}$ nonterminals, each of which can expand to any word whatsoever. In practice Word_t will expand to those words most strongly associated with topic t, while $\mathrm{Word}_{\mathrm{None}}$ will expand to those words not associated with any topic.

[1] In fact, the unigram grammar is equivalent to a HMM, but the PCFG parameterisation makes clear the relationship between grounded learning and estimation of grammar rule weights.

$$
\begin{array}{ll}
\text{Sentence} \rightarrow \text{Topic}_t\ \text{Collocs}_t & \forall t \in T' \\
\text{Collocs}_t \rightarrow \text{Colloc}_t\ (\text{Collocs}_t) & \forall t \in T' \\
\text{Collocs}_t \rightarrow \text{Colloc}_{\text{None}}\ (\text{Collocs}_t) & \forall t \in T \\
\underline{\text{Colloc}_t} \rightarrow \text{Words}_t & \forall t \in T' \\
\text{Words}_t \rightarrow \text{Word}_t\ (\text{Words}_t) & \forall t \in T' \\
\text{Words}_t \rightarrow \text{Word}_{\text{None}}\ (\text{Words}_t) & \forall t \in T \\
\underline{\text{Word}_t} \rightarrow \text{Word} & \forall t \in T' \\
\text{Word} \rightarrow w & \forall w \in W
\end{array}
$$

Figure 5: The rule schema that generate the collocation adaptor grammar. Adapted nonterminals are indicated via underlining. Here T is the set of all non-None available topics, $T' = T \cup \{\text{None}\}$, and W is the set of words appearing in the utterances. The rules expanding the Topic_t nonterminals are exactly as in unigram PCFG.

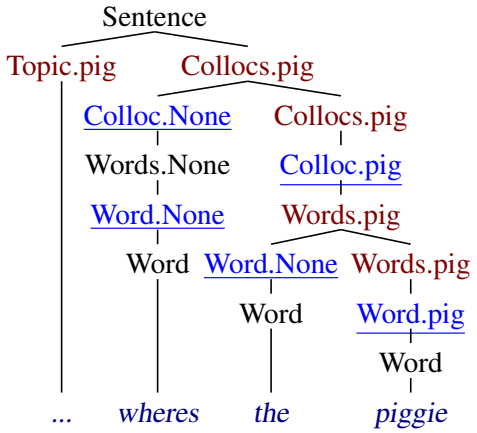

Figure 6: Sample parse generated by the collocation adaptor grammar. The adapted nonterminals Colloc_t and Word_t are shown underlined; the subtrees they dominate are "cached" by the adaptor grammar. The prefix (not shown here) is parsed exactly as in the Unigram PCFG.

2.2 Adaptor grammars

Our other grounded learning models are based on reductions of grounded learning to adaptor grammar inference problems. Adaptor grammars are a framework for stating a variety of Bayesian nonparametric models defined in terms of a hierarchy of Pitman-Yor Processes: see Johnson et al. (2007) for a formal description. Informally, an adaptor grammar is specified by a set of rules just as in a PCFG, plus a set of *adapted nonterminals*. The set of trees generated by an adaptor grammar is the same as the set of trees generated by a PCFG with the same rules, but the generative process differs. Non-adapted nonterminals in an adaptor grammar expand just as they do in a PCFG: the probability of choosing a rule is specified by its probability. However, the expansion of an adapted nonterminal depends on how it expanded in previous derivations. An adapted nonterminal can directly expand to a subtree with probability proportional to the number of times that subtree has been previously generated; it can also "back off" to expand using a grammar rule, just as in a PCFG, with probability proportional to a constant.[2]

Thus an adaptor grammar can be viewed as caching each tree generated by each adapted nonterminal, and regenerating it with probability proportional to the number of times it was previously generated (with some probability mass reserved to generate "new" trees). This enables adaptor grammars to generalise over subtrees of arbitrary size. Generic software is available for adaptor grammar inference, based either on Variational Bayes (Cohen et al., 2010) or Markov Chain Monte Carlo (Johnson and Goldwater, 2009). We used the latter software because it is capable of performing hyper-parameter inference for the PCFG rule probabilities and the Pitman-Yor Process parameters. We used the "out-of-the-box" settings for this software, i.e., uniform priors on all PCFG rule parameters, a $\text{Beta}(2, 1)$ prior on the Pitman-Yor a parameters and a "vague" $\text{Gamma}(100, 0.01)$ prior on the Pitman-Yor b parameters. (Presumably performance could be improved if the priors were tuned, but we did not explore this here).

Here we explore a simple "collocation" extension to the unigram PCFG which associates multiword collocations, rather than individual words, with topics. Hardisty et al. (2010) showed that this significantly improved performance in a sentiment analysis task.

The collocation adaptor grammar in Figure 5 generates the words of the utterance as a sequence of collocations, each of which is a sequence of words. Each collocation is either associated with the sentence topic or with the None topic, just like words in the unigram model. Figure 6 shows a sample parse generated by the collocation adaptor grammar.

We also experimented with a variant of the unigram and collocation grammars in which the topic-specific word distributions Word_t for each $t \in T$

[2]This is a description of Chinese Restaurant Processes, which are the predictive distributions for Dirichlet Processes. Our adaptor grammars are actually based on the more general Pitman-Yor Processes, as described in Johnson and Goldwater (2009).

Model	Social cues	Utterance topic				Word topic			Lexicon		
		acc.	f-score	prec.	rec.	f-score	prec.	rec.	f-score	prec.	rec.
unigram	none	0.3395	0.4044	0.3249	0.5353	0.2007	0.1207	0.5956	0.1037	0.05682	0.5952
unigram	all	0.4907	0.6064	0.4867	**0.8043**	0.295	0.1763	**0.9031**	0.1483	0.08096	**0.881**
colloc	none	0.4331	0.3513	0.3272	0.3792	0.2431	0.1603	0.5028	0.08808	0.04942	0.4048
colloc	all	0.5837	0.598	0.5623	0.6384	**0.4098**	**0.2702**	0.8475	0.1671	0.09422	0.7381
unigram′	none	0.3261	0.3767	0.3054	0.4914	0.1893	0.1131	0.5811	0.1167	0.06583	0.5122
unigram′	all	0.5117	**0.6106**	0.4986	0.7875	0.2846	0.1693	0.891	0.1684	0.09402	0.8049
colloc′	none	0.5238	0.3419	0.3844	0.3078	0.2551	0.1732	0.4843	0.2162	0.1495	0.3902
colloc′	all	**0.6492**	0.6034	**0.6664**	0.5514	0.3981	0.2613	0.8354	**0.3375**	**0.2269**	0.6585

Figure 7: Utterance topic, word topic and lexicon results for all models, on data with and without social cues. The results for the variant models, in which Word_t nonterminals expand via $\mathrm{Word}_{\mathrm{None}}$, are shown under unigram′ and colloc′. Utterance topic shows how well the model discovered the intended topics at the utterance level, word topic shows how well the model associates word tokens with topics, and lexicon shows how well the topic most frequently associated with a word type matches an external word-topic dictionary. In this figure and below, "colloc" abbreviates "collocation", "acc." abbreviates "accuracy", "prec." abbreviates "precision" and "rec." abbreviates "recall".

(the set of non-None available topics) expand via $\mathrm{Word}_{\mathrm{None}}$ non-terminals. That is, in the variant grammars topical words are generated with the following rule schema:

$$\begin{aligned} \mathrm{Word}_t &\rightarrow \mathrm{Word}_{\mathrm{None}} && \forall t \in T \\ \mathrm{Word}_{\mathrm{None}} &\rightarrow \mathrm{Word} && \\ \mathrm{Word} &\rightarrow w && \forall w \in W \end{aligned}$$

In these variant grammars, the $\mathrm{Word}_{\mathrm{None}}$ nonterminal generates all the words of the language, so it defines a generic "background" distribution over all the words, rather than just the nontopical words. An effect of this is that the variant grammars tend to identify fewer words as topical.

3 Experimental evaluation

We performed grammatical inference using the adaptor grammar software described in Johnson and Goldwater (2009).[3] All experiments involved 4 runs of 5,000 samples each, of which the first 2,500 were discarded for "burn-in".[4] From these samples we extracted the modal (i.e., most frequent) analysis, which we evaluated as described below. The results of evaluating each model on the corpus with social cues, and on another corpus identical except that the social cues have been removed, are presented in Figure 7.

Each model was evaluated on each corpus as follows. First, we extracted the utterance's topic from the modal parse (this can be read off the Topic_t nodes), and compared this to the intended topics annotated in the corpus. The frequency with which the models' predicted topics exactly matches the intended topics is given under "utterance topic accuracy"; the f-score, precision and recall of each model's topic predictions are also given in the table.

Because our models all associate word tokens with topics, we can also evaluate the accuracy with which word tokens are associated with topics. We constructed a small dictionary which identifies the words that can be used as the head of a phrase to refer to the topical objects (e.g., the dictionary indicates that *dog*, *doggie* and *puppy* name the topical object DOG). Our dictionary is relatively conservative; between one and eight words are associated with each topic. We scored the topic label on each word token in our corpus as follows. A topic label is scored as correct if it is given in our dictionary and the topic is one of the intended topics for the utterance. The "word topic" entries in Figure 7 give the results of this evaluation.

[3] Because adaptor grammars are a generalisation of PCFGs, we could use the adaptor grammar software to estimate the unigram model.

[4] We made no effort to optimise the computation, but it seems the samplers actually stabilised after around a hundred iterations, so it was probably not necessary to sample so extensively. We estimated the error in our results by running our most complex model (the colloc′ model with all social cues) 20 times (i.e., 20×8 chains for 5,000 iterations) so we could compute the variance of each of the evaluation scores (it is reasonable to assume that the simpler models will have smaller variance). The standard deviation of all utterance topic and word topic measures is between 0.005 and 0.01; the standard deviation for lexicon f-score is 0.02, lexicon precision is 0.01 and lexicon recall is 0.03. The adaptor grammar software uses a sentence-wise blocked sampler, so it requires fewer iterations than a pointwise sampler. We used 5,000 iterations because this is the software's default setting; evaluating the trace output suggests it only takes several hundred iterations to "burn in". However, we ran 8 chains for 25,000 iterations of the colloc′ model; as expected the results of this run are within two standard deviations of the results reported above.

Model	Social cues	Utterance topic				Word topic			Lexicon		
		acc.	f-score	prec.	rec.	f-score	prec.	rec.	f-score	prec.	rec.
unigram	none	0.3395	0.4044	0.3249	0.5353	0.2007	0.1207	0.5956	0.1037	0.05682	0.5952
unigram	+*child.eyes*	**0.4573**	**0.5725**	**0.4559**	**0.7694**	**0.2891**	**0.1724**	**0.8951**	**0.1362**	**0.07415**	**0.8333**
unigram	+*child.hands*	0.3399	0.4011	0.3246	0.5247	0.2008	0.121	0.5892	0.09705	0.05324	0.5476
unigram	+*mom.eyes*	0.338	0.4023	0.3234	0.5322	0.1992	0.1198	0.5908	0.09664	0.053	0.5476
unigram	+*mom.hands*	0.3563	0.4279	0.3437	0.5667	0.1984	0.1191	0.5948	0.09959	0.05455	0.5714
unigram	+*mom.point*	0.3063	0.3548	0.285	0.4698	0.1806	0.1086	0.5359	0.09224	0.05057	0.5238
colloc	none	0.4331	0.3513	0.3272	0.3792	0.2431	0.1603	0.5028	0.08808	0.04942	0.4048
colloc	+*child.eyes*	**0.5159**	**0.5006**	**0.4652**	**0.542**	**0.351**	**0.2309**	**0.7312**	**0.1432**	**0.07989**	**0.6905**
colloc	+*child.hands*	0.4827	0.4275	0.3999	0.4592	0.2897	0.1913	0.5964	0.1192	0.06686	0.5476
colloc	+*mom.eyes*	0.4697	0.4171	0.3869	0.4525	0.2708	0.1781	0.5642	0.1013	0.05666	0.4762
colloc	+*mom.hands*	0.4747	0.4251	0.3942	0.4612	0.274	0.1806	0.5666	0.09548	0.05337	0.4524
colloc	+*mom.point*	0.4228	0.3378	0.3151	0.3639	0.2575	0.1716	0.5157	0.09278	0.05202	0.4286

Figure 8: Effect of using just one social cue on the experimental results for the unigram and collocation models. The "importance" of a social cue can be quantified by the degree to which the model's evaluation score improves when using a corpus containing that social cue relative to its evaluation score when using a corpus without any social cues. The most important social cue is the one which causes performance to improve the most.

Finally, we extracted a lexicon from the parsed corpus produced by each model. We counted how often each word type was associated with each topic in our sampler's output (including the None topic), and assigned the word to its most frequent topic. The "lexicon" entries in Figure 7 show how well the entries in these lexicons match the entries in the manually-constructed dictionary discussed above.

There are 10 different evaluation scores, and no model dominates in all of them. However, the top-scoring result in every evaluation is always for a model trained using social cues, demonstrating the importance of these social cues. The variant collocation model (trained on data with social cues) was the top-scoring model on four evaluation scores, which is more than any other model.

One striking thing about this evaluation is that the recall scores are all much higher than the precision scores, for each evaluation. This indicates that all of the models, especially the unigram model, are labelling too many words as topical. This is perhaps not too surprising: because our models completely lack any notion of syntactic structure and simply model the association between words and topics, they label many non-nouns with topics (e.g., *woof* is typically labelled with the topic DOG).

3.1 Evaluating the importance of social cues

It is scientifically interesting to be able to evaluate the importance of each of the social cues to grounded learning. One way to do this is to study the effect of adding or removing social cues from the corpus on the ability of our models to perform grounded learning. An important social cue should have a large impact on our models' performance; an unimportant cue should have little or no impact.

Figure 8 compares the performance of the unigram and collocation models on corpora containing a single social cue to their performance on the corpus without any social cues, while Figure 9 compares the performance of these models on corpora containing all but one social cue to the corpus containing all of the social cues. In both of these evaluations, with respect to all 10 evaluation measures, the *child.eyes* social cue had the most impact on model performance.

Why would the child's own gaze be more important than the caregiver's? Perhaps caregivers are *following in*, i.e., talking about objects that their children are interested in (Baldwin, 1991). However, another possible explanation is that this result is due to the general continuity of conversational topics over time. Frank et al. (to appear) show that for the current corpus, the topic of the preceding utterance is very likely to be the topic of the current one also. Thus, the child's eyes might be a good predictor because they reflect the fact that the child's attention has been drawn to an object by previous utterances.

Notice that these two possible explanations of the importance of the *child.eyes* cue are diametrically opposed; the first explanation claims that the cue is important because the child is driving the discourse, while the second explanation claims that the cue is important because the child's gaze follows the topic of the caregiver's previous utterance. This sort of question about causal relationships in conversations may be very difficult to answer using standard descriptive techniques, but it may be an interesting av-

Model	Social cues	Utterance topic				Word topic			Lexicon		
		acc.	f-score	prec.	rec.	f-score	prec.	rec.	f-score	prec.	rec.
unigram	all	0.4907	0.6064	0.4867	0.8043	0.295	0.1763	0.9031	0.1483	0.08096	0.881
unigram	−*child.eyes*	**0.3836**	**0.4659**	**0.3738**	**0.6184**	**0.2149**	**0.1286**	**0.6546**	**0.1111**	**0.06089**	**0.6341**
unigram	−*child.hands*	0.4907	0.6063	0.4863	0.8051	0.296	0.1769	0.9056	0.1525	0.08353	0.878
unigram	−*mom.eyes*	0.4799	0.5974	0.4768	0.7996	0.2898	0.1727	0.9007	0.1551	0.08486	0.9024
unigram	−*mom.hands*	0.4871	0.5996	0.4815	0.7945	0.2925	0.1746	0.8991	0.1561	0.08545	0.9024
unigram	−*mom.point*	0.4875	0.6033	0.4841	0.8004	0.2934	0.1752	0.9007	0.1558	0.08525	0.9024
colloc	all	0.5837	0.598	0.5623	0.6384	0.4098	0.2702	0.8475	0.1671	0.09422	0.738
colloc	−*child.eyes*	**0.5604**	**0.5746**	**0.529**	**0.6286**	**0.39**	**0.2561**	**0.8176**	**0.1534**	**0.08642**	**0.6829**
colloc	−*child.hands*	0.5849	0.6	0.5609	0.6451	0.4145	0.273	0.8612	0.1662	0.09375	0.7317
colloc	−*mom.eyes*	0.5709	0.5829	0.5457	0.6255	0.4036	0.2655	0.8418	0.1662	0.09375	0.7317
colloc	−*mom.hands*	0.5795	0.5935	0.5571	0.6349	0.4038	0.2653	0.8442	0.1788	0.1009	0.7805
colloc	−*mom.point*	0.5851	0.6006	0.5607	0.6467	0.4097	0.2685	0.8644	0.1742	0.09841	0.7561

Figure 9: Effect of using all but one social cue on the experimental results for the unigram and collocation models. The "importance" of a social cue can be quantified by the degree to which the model's evaluation score degrades when that just social cue is removed from the corpus, relative to its evaluation score when using a corpus without all social cues. The most important social cue is the one which causes performance to degrade the most.

enue for future investigation using more structured models such as those proposed here.[5]

4 Conclusion and future work

This paper presented four different grounded learning models that exploit social cues. These models are all expressed via reductions to grammatical inference problems, so standard "off the shelf" grammatical inference tools can be used to learn them. Here we used the same adaptor grammar software tools to learn all these models, so we can be relatively certain that any differences we observe are due to differences in the models, rather than quirks in the software.

Because the adaptor grammar software performs full Bayesian inference, including for model parameters, an unusual feature of our models is that we did not need to perform any parameter tuning whatsoever. This feature is particularly interesting with respect to the parameters on social cues. Psychological proposals have suggested that children may discover that particular social cues help in establishing reference (Baldwin, 1993; Hollich et al., 2000), but prior modeling work has often assumed that cues, cue weights, or both are prespecified. In contrast, the models described here could in principle discover a wide range of different social conventions.

[5]A reviewer suggested that we can test whether *child.eyes* effectively provides the same information as the previous topic by adding the previous topic as a (pseudo-) social cue. We tried this, and *child.eyes* and *previous.topic* do in fact seem to convey very similar information: e.g., the model with *previous.topic* and without *child.eyes* scores essentially the same as the model with all social cues.

Our work instantiates the strategy of investigating the structure of children's learning environment using "ideal learner" models. We used our models to investigate scientific questions about the role of social cues in grounded language learning. Because the performance of all four models studied in this paper improve dramatically when provided with social cues in all ten evaluation metrics, this paper provides strong support for the view that social cues are a crucial information source for grounded language learning.

We also showed that the importance of the different social cues in grounded language learning can be evaluated using "add one cue" and "subtract one cue" methodologies. According to both of these, the *child.eyes* cue is the most important of the five social cues studied here. There are at least two possible reasons for this: the caregiver's topic could be determined by the child's gaze, or the *child.eyes* cue could be providing our models with information about the topic of the previous utterance.

Incorporating topic continuity and anaphoric dependencies into our models would be likely to improve performance. This improvement might also help us distinguish the two hypotheses about the *child.eyes* cue. If the *child.eyes* cue is just providing indirect information about topic continuity, then the importance of the *child.eyes* cue should decrease when we incorporate topic continuity into our models. But if the child's gaze is in fact determining the care-giver's topic, then *child.eyes* should remain a strong cue even when anaphoric dependencies and topic continuity are incorporated into our models.

Acknowledgements

This research was supported under the Australian Research Council's *Discovery Projects* funding scheme (project number DP110102506).

References

Dare A. Baldwin. 1991. Infants' contribution to the achievement of joint reference. *Child Development*, 62(5):874–890.

Dare A. Baldwin. 1993. Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20:395–395.

Benjamin Börschinger, Bevan K. Jones, and Mark Johnson. 2011. Reducing grounded learning tasks to grammatical inference. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1416–1425, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

M. Carpenter, K. Nagell, M. Tomasello, G. Butterworth, and C. Moore. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*.

E.V. Clark. 1987. The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1:33.

Shay B. Cohen, David M. Blei, and Noah A. Smith. 2010. Variational inference for adaptor grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 564–572, Los Angeles, California, June. Association for Computational Linguistics.

Michael Frank, Noah Goodman, and Joshua Tenenbaum. 2008. A Bayesian framework for cross-situational word-learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 457–464, Cambridge, MA. MIT Press.

Michael C. Frank, Joshua Tenenbaum, and Anne Fernald. to appear. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*.

Eric A. Hardisty, Jordan Boyd-Graber, and Philip Resnik. 2010. Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 284–292, Stroudsburg, PA, USA. Association for Computational Linguistics.

G.J. Hollich, K. Hirsh-Pasek, and R. Golinkoff. 2000. Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, June. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Mark Johnson, Katherine Demuth, Michael Frank, and Bevan Jones. 2010. Synergies in learning words and their referents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1018–1026.

Mark Johnson. 2008. Using adaptor grammars to identifying synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Mark Johnson. 2010. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July. Association for Computational Linguistics.

Patricia K. Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences USA*, 100(15):9096–9101.

Jeffrey Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

L.B. Smith, S.S. Jones, B. Landau, L. Gershkoff-Stowe, and L. Samuelson. 2002. Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1):13.

Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.