Parsing entire discourses as very long strings: Capturing topic continuity in grounded language learning

Minh-Thang Luong
Department of Computer Science
Stanford University
Stanford, California
lmthang@stanford.edu

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

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

Abstract

Grounded language learning, the task of mapping from natural language to a representation of meaning, has attracted more and more interest in recent years. In most work on this topic, however, utterances in a conversation are treated independently and discourse structure information is largely ignored. In the context of language acquisition, this independence assumption discards cues that are important to the learner, e.g., the fact that consecutive utterances are likely to share the same referent (Frank et al., 2013). The current paper describes an approach to the problem of simultaneously modeling grounded language at the sentence and discourse levels. We combine ideas from parsing and grammar induction to produce a parser that can handle long input strings with thousands of tokens, creating parse trees that represent full discourses. By casting grounded language learning as a grammatical inference task, we use our parser to extend the work of Johnson et al. (2012), investigating the importance of discourse continuity in children's language acquisition and its interaction with social cues. Our model boosts performance in a language acquisition task and yields good discourse segmentations compared with human annotators.

1 Introduction

Learning mappings between natural language (NL) and meaning representations (MR) is an important goal for both computational linguistics and cognitive science. Accurately learning novel mappings is crucial in grounded language understanding tasks and such systems can suggest insights into the nature of children language learning.

Two influential examples of grounded language learning tasks are the sportscasting task, RoboCup, where the NL is the set of running commentary and the MR is the set of logical forms representing actions like kicking or passing (Chen and Mooney, 2008), and the cross-situational word-learning task, where the NL is the caregiver's utterances and the MR is the set of objects present in the context (Siskind, 1996; Yu and Ballard, 2007). Work in these domains suggests that, based on the co-occurrence between words and their referents in context, it is possible to learn mappings between NL and MR even under substantial ambiguity.

Nevertheless, contexts like RoboCup—where every single utterance is grounded—are extremely rare. Much more common are cases where a single topic is introduced and then discussed at length throughout a discourse. In a television news show, for example, a topic might be introduced by presenting a relevant picture or video clip. Once the topic is introduced, the anchors can discuss it by name or even using a pronoun without showing a picture. The *discourse* is grounded without having to ground every utterance.

Moreover, although previous work has largely treated utterance order as independent, the order of utterances is critical in grounded discourse contexts: if the order is scrambled, it can become impossible to recover the topic. Supporting this idea, Frank et al. (2013) found that topic continuity—the tendency to talk about the same topic in multiple utterances that are contiguous in time—is both prevalent and informative for word learning. This paper examines the importance of topic continuity through a grammatical inference problem. We build on Johnson et al. (2012)'s work that used grammatical inference to

Transactions of the Association for Computational Linguistics, 1 (2013) 315–326. Action Editor: Mark Steedman.
Submitted 2/2013; Revised 6/2013; Published 7/2013.

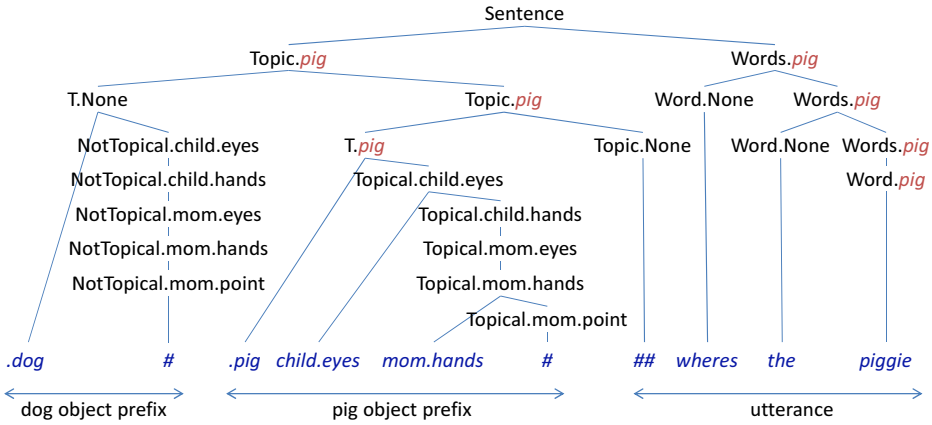

Figure 1: **Unigram Social Cue PCFGs** (Johnson et al., 2012) – shown is a parse tree of the input utterance "wheres the piggie" accompanied with social cue prefixes, indicating that the caregiver is holding a pig toy while the child is looking at it; at the same time, a dog toy is present in the screen.

learn word-object mappings and to investigate the role of social information (cues like eye-gaze and pointing) in a child language acquisition task.

Our main contribution lies in the novel integration of existing techniques and algorithms in parsing and grammar induction to offer a complete solution for simultaneously modeling grounded language at the sentence and discourse levels. Specifically, we: (1) use the Earley algorithm to exploit the special structure of our grammars, which are deterministic or have at most bounded ambiguity, to achieve approximately linear parsing time; (2) suggest a rescaling approach that enables us to build a PCFG parser capable of handling very long strings with thousands of tokens; and (3) employ Variational Bayes for grammatical inference to obtain better grammars than those given by the EM algorithm.

By parsing entire discourses at once, we shed light on a scientifically interesting question about why the child's own gaze is a positive cue for word learning (Johnson et al., 2012). Our data provide support for the hypothesis (from previous work) that caregivers "follow in": they name objects that the child is already looking at (Tomasello and Farrar, 1986). In addition, our discourse model produces a performance improvement in a language acquisition task and yields good discourse segmentations compared with human annotators.

2 Related Work

Supervised semantic parsers. Previous work has developed supervised semantic parsers to map sentences to meaning representations of various forms, including meaning hierarchies (Lu et al., 2008) and, most dominantly, λ-calculus expressions (Zettlemoyer and Collins, 2005; Zettlemoyer, 2007; Wong and Mooney, 2007; Kwiatkowski et al., 2010). These approaches rely on training data of annotated sentence-meaning pairs, however. Such data are costly to obtain and are quite different from the experience of language learners.

Grounded Language Learning. In contrast to semantic parsers, grounded language learning systems aim to learn the meanings of words and sentences given an observed world state (Yu and Ballard, 2004; Gorniak and Roy, 2007). A growing body of work in this field employs distinct techniques from a wide variety of perspectives from *text-to-record* alignment using structured classification (Barzilay and Lapata, 2005; Snyder and Barzilay, 2007), iterative retraining (Chen et al., 2010), and generative models of segmentation and alignment (Liang et al., 2009) to *text-to-interaction* mapping using reinforcement learning (Branavan et al., 2009; Vogel and Jurafsky, 2010), graphical model semantics representation (Tellex et al., 2011a; Tellex et al., 2011b), and Combinatory Categorial Grammar (Artzi and Zettlemoyer, 2013). A number of systems have also used alternative forms of supervision, including sentences paired with responses (Clarke et al., 2010; Goldwasser and Roth, 2011; Liang et al., 2011) and no supervision (Poon and Domingos, 2009; Gold-

wasser et al., 2011).

Recent work has also introduced an alternative approach to grounded learning by reducing it to a grammatical inference problem. Börschinger et al. (2011) casted the problem of learning a semantic parser as a PCFG induction task, achieving state-of the art performance in the RoboCup domain. Kim and Mooney (2012) extended the technique to make it tractable for more complex problems. Later, Kim and Mooney (2013) adapted discriminative reranking to the grounded learning problem using a form of weak supervision. We employ this general grammatical inference approach in the current work.

Children Language Acquisition. In the context of language acquisition, Frank et al. (2008) proposed a system that learned words and jointly inferred speakers' intended referent (utterance topic) using graphical models. Johnson et al. (2012) used grammatical inference to demonstrate the importance of social cues in children's early word learning. We extend this body of work by capturing discourse-based dependencies among utterances rather than treating each utterance independently.

Discourse Parsing. A substantial literature has examined formal representations of discourse across a wide variety of theoretical perspectives (Mann and Thompson, 1988; Scha and Polanyi, 1988; Hobbs, 1990; Lascarides and Asher, 1993; Knott and Sanders, 1997). Although much of this work was highly influential, Marcu (1997)'s work on discourse parsing brought this task to special prominence. Since then, more and more sophisticated models of discourse analysis have been developed:, e.g., (Marcu, 1999; Soricut and Marcu, 2003; Forbes et al., 2003; Polanyi et al., 2004; Baldridge and Lascarides, 2005; Subba and Di Eugenio, 2009; Hernault et al., 2010; Lin et al., 2012; Feng and Hirst, 2012). Our contribution to work on this task is to examine latent discourse structure specifically in grounded language learning.

3 A Grounded Learning Task

Our focus in this paper is to develop computational models that help us better understand children's language acquisition. The goal is to learn both the long term lexicon of mappings between words and objects (language learning) as well as the intended topic of individual utterances (language comprehension). We consider a corpus of child-directed speech annotated with social cues, described in (Frank et al., 2013). There are a total of 4,763 utterances in the corpus, each of which is orthographically-transcribed from videos of caregivers playing with pre-linguistic children of various ages (6, 12, and 18 months) during home visits.[1] Each utterance was hand-annotated with objects present in the (non-linguistic) context, e.g. `dog` and `pig` (Figure 1), together with sets of social cues, one set per object. The social cues describe objects the care-giver is looking at (`mom.eyes`), holding onto (`mom.hands`), or pointing to (`mom.point`); similarly, for (`child.eyes`) and (`child.hands`).

3.1 Sentence-level Models

Motivated by the importance of social information in children's early language acquisition (Carpenter et al., 1998), Johnson et al. (2012) proposed a joint model of *non-linguistic information* including the physical context and social cues, and the *linguistic content* of individual utterances. They framed the joint inference problem of inferring word-object mappings and inferring sentence topics as a grammar induction task where input strings are utterances prefixed with non-linguistic information. Objects present in the non-linguistic context of an utterance are considered its potential topics. There is also a special null topic, `None`, to indicate non-topical utterances. The goal of the model is then to select the most probable topic for each utterance.

Top-level rules, $\texttt{Sentence} \rightarrow \texttt{Topic}_t\ \texttt{Words}_t$ (unigram PCFG) or $\texttt{Sentence} \rightarrow \texttt{Topic}_t\ \texttt{Collocs}_t$ (collocation Adaptor Grammar), are tailored to link the two modalities (t ranges over T', the set of all available topics (T) and `None`). These rules enforce sharing of topics between prefixes ($\texttt{Topic}_t$) and words ($\texttt{Words}_t$ or $\texttt{Collocs}_t$). Each word in the utterance is drawn from either a topic-specific distribution $\texttt{Word}_t$ or a general "null" distribution $\texttt{Word}_{None}$.

As illustrated in Figure 1, the selected topic, `pig`, is propagated down to the input string through two paths: (a) through *topical* nodes until an object is

[1] Caregivers were given pairs of toys to play with, e.g. a stuffed dog and pig, or a wooden car and truck.

reached, in this case the *.pig* object, and (b) through *lexical* nodes to topical word tokens, e.g. *piggie*. Social cues are then generated by a series of binary decisions as detailed in Johnson et al. (2012). The key feature of these grammars is that parameter inference corresponds both to learning word-topic relations and learning the salience of social cues in grounded learning.

In the current work, we restrict our attention to only the unigram PCFG model to focus on investigating the role of topic continuity. Unlike the approach of Johnson et al. (2012), which uses Markov Chain Monte Carlo techniques to perform grammatical inference, we experiment with *Variational Bayes* methods, detailed in Section 6.

3.2 A Discourse-level Model

Topic continuity—the tendency to group utterances into coherent discourses about a single topic—may be an important source of information for children learning the meanings of words (Frank et al., 2013). To address this issue, we consider a new discourse-level model of grounded language that captures dependencies between utterances. By linking multiple utterances in a single parse, our proposed grammatical formalism is a bigram Markov process that models transitions among utterance topics.

Our grammar starts with a root symbol `Discourse`, which then selects a starting topic through a set of discourse *initial* rules, $\texttt{Discourse} \rightarrow \texttt{Discourse}_t$ for $t \in T'$. Each of the $\texttt{Discourse}_t$ nodes generates an utterance of the same topic, and advances into other topics through *transition* rules, $\texttt{Discourse}_t \rightarrow \texttt{Sentence}_t\ \texttt{Discourse}_{t'}$ for $t' \in T'$. Discourses terminate by *ending* rules, $\texttt{Discourse}_t \rightarrow \texttt{Sentence}_t$. Other rules in the unigram PCFG model by Johnson are reused except for the top-level rules in which we replace the non-terminal `Sentence` by topic-specific ones $\texttt{Sentence}_t$.

3.3 Parsing Discourses and Challenges

Using a discourse-level grammar, we must parse a concatenation of all the utterances (with annotations) in each conversation. This concatenation results in an extremely long string: in the social-cue corpus (Frank et al., 2013), the average length of these per-recording concatenations is 2152 tokens (σ=972). Parsing such strings poses many challenges for existing algorithms.

For familiar algorithms such as CYK, runtime quickly becomes enormous: the time complexity of CYK is $O(n^3)$ for an input of length n. Fortunately, we can take advantage of a special structural property of our grammars. The shape of the parse tree is completely determined by the input string; the only variation is in the topic annotations in the nonterminal labels. So even though the number of possible parses grows exponentially with input length n, the number of possible constituents grows only linearly with input length, and the possible constituents can be identified from the left context.[2] These constraints ensure that the Earley algorithm[3] (Earley, 1970) will parse an input of length n with this grammar in time $O(n)$.

A second challenge in parsing very long strings is that the probability of a parse is the product of the probabilities of the rules involved in its derivation. As the length of a derivation grows linearly with the length of the input, the parse probabilities decrease exponentially as a function of sentence length, causing floating-point underflow on inputs of even moderate length. The standard method for handling this is to compute log probabilities (which decrease linearly as a function of input length, rather than exponentially), but as we explain later (Section 5), we can use the ability of the Earley algorithm to compute prefix probabilities (Stolcke, 1995) to rescale the probability of the parse incrementally and avoid floating-point underflows.

In the next section, we provide background information on the Earley algorithm for PCFGs, the prefix probability scheme we use, and the inside-outside algorithm in the Earley context.

4 Background

4.1 Earley Algorithm for PCFGs

The Earley algorithm was developed by Earley (1970) and known to be efficient for certain kinds of CFGs (Aho and Ullman, 1972). An Earley parser

[2]The prefix markers # and ## and the topic markers such as ".dog" enable a left-to- right parser to unambiguously identify its location in the input string.

[3]In order to achieve linear time the parsing chart must have suitable indexing; see Aho and Ullman (1972), Leo (1991) and Aycock and Horspool (2002) for details.

constructs left-most derivations of strings, using dotted productions to keep track of partial derivations. Specifically, each state in an Earley parser is represented as $[l, r]$: $X \rightarrow \alpha \,.\, \beta$ to indicate that input symbols $x_l, \ldots, x_{r-1}$ have been processed and the parser is expecting to expand β. States are generated on the fly using three transition operations: *predict* (add states to charts), *scan* (shift dots across terminals), and *complete* (merge two states). Figure 2 shows an example of a completion step which also illustrates the implicit binarization automatically done in Earley algorithm.

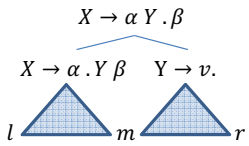

Figure 2: **Completion step** – merging two states $[l, m]$: $X \rightarrow \alpha \,.\, Y\beta$ and $[m, r]$:$Y \rightarrow \nu \,.$ to produce a new state $[l, r]$: $X \rightarrow \alpha Y \,.\, \beta$.

In order to handle PCFGs, Stolcke (1995) extends the Earley parsing algorithm to introduce the notion of an *Earley path* being a sequence of states linked by Earley operations. By establishing a one-to-one mapping between partial derivations and Earley paths, Stolcke could then assign each path a derivation probability, that is the product of the all rule probabilities used in the predicted states of that path. Here, each production $X \rightarrow \nu$ corresponds to a predicted state $[l, l] : X \rightarrow .\, \nu$.

Besides parsing, being able to compute string and prefix probabilities by summing derivation probabilities is also of great importance. To compute these sums efficiently, each Earley state is attached with a *forward* and an *inner* probability which are updated incrementally as new states are spawned by the three transition operations.

4.2 Forward and Prefix Probabilities

Intuitively, the forward probability of a state $[l, r]$: $X \rightarrow \alpha \,.\, \beta$ is the probability of an Earley path through that state, generating input up to position r-1. This probability generalizes a similar concept in HMM and lends itself to the computation of prefix probabilities, sums of forward probabilities over scanned states yielding a prefix x.

Computing prefix probabilities is important because it enables probabilistic prediction of possible follow-words x_{i+1} as $P(x_{i+1}|x_0 \ldots x_i) = \frac{P(x_0 \ldots x_i x_{i+1})}{P(x_0 \ldots x_i)}$ (Jelinek and Lafferty, 1991). These conditional probabilities allow estimation of the incremental costs of a stack decoder (Bahl et al., 1983). In (Huang and Sagae, 2010), a conceptually similar prefix cost is defined to order states in a beam search decoder. Moreover, the negative logarithm of such conditional probabilities are termed as *surprisal* values in the psycholinguistics literature (e.g., Hale, 2001; Levy, 2008), to describe how difficult a word is in a given context. Interestingly, we show that prefix probabilities lead us to construct a parser that could parse extremely long strings next.

4.3 Inside Outside Algorithm

To extend the Inside Outside (IO) algorithm (Baker, 1979) to the Earley context, Stolcke introduced inner and outer probabilities which generalize the inside and outside probabilities in the IO algorithm. Specifically, the inner probability of a state $[l, r]$: $X \rightarrow \alpha \,.\, \beta$ is the probability of generating an input substring $x_l, \ldots, x_{r-1}$ from a non-terminal X using a production $X \rightarrow \alpha\ \beta$.[4]

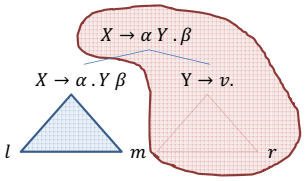

Figure 3: **Inner and outer probabilities.** The outer probability of $X \rightarrow \alpha \,.\, Y\beta$ is a sum of all products of its parent outer probability ($X \rightarrow \alpha Y \,.\, \beta$) and its sibling inner probability ($Y \rightarrow \nu \,.$). Similarly, the outer probability of $Y \rightarrow \nu \,.$ is derived from the outer probability of $X \rightarrow \alpha Y \,.\, \beta$ and the inner probability of $X \rightarrow \alpha \,.\, Y\beta$.

Once all inner probabilities have been populated in a forward pass, *outer* probabilities are derived backward, starting from the outer probability of the goal state $[0, n] :\rightarrow S \,.$ being 1. Here, each Earley state is associated with an outer probability which complements the inner probability by referring precisely to those parts (not covered by the corresponding inner probability) of the complete paths generating the input string x. The implicit binarization in

[4]Summing up inner probabilities of all states $Y \rightarrow \nu \,.$ exactly yields Baker's inside probability for Y.

Earley parsing allows outer probabilities to be accumulated in a similar way as its counterpart in the IO algorithm (see Figure 3).

These quantities allow for efficient grammatical inference in which the expected count of each rule $X \rightarrow \lambda$ given a string x is computed as:

$$c(X \rightarrow \lambda | x) = \frac{\sum_{s:[l,r]X \rightarrow .\lambda} \text{outer}(s) \cdot \text{inner}(s)}{P(S \Rightarrow^* x)}. \quad (1)$$

5 A Rescaling Approach for Parsing

Our parser originated from the prefix probability parser by Levy (2008), but has diverged markedly since then. The parser, called Earleyx[5], is capable of producing Viterbi parses and performing grammatical induction based on the expectation-maximization and variational Bayes algorithms.

To tackle the underflow problem posed when parsing discourses (§3.3), we borrow the rescaling concept from HMMs (Rabiner, 1990) to extend the probabilistic Earley algorithm. Specifically, the probability of each Earley path is scaled by a constant c_i each time it passes through a scanned state generating the input symbol x_i. In fact, each path passes through each scanned state exactly once, so we consistently accumulate scaling factors for the forward and inner probabilities of a state $[l, r] : X \rightarrow \alpha \, . \, \beta$ as $c_0 \ldots c_{r-1}$ and $c_l \ldots c_{r-1}$ respectively.

Arguably, the most intuitive choice of the scaling factors are the prefix probabilities, which essentially resets the probability of any Earley path starting from any position i to 1. Concretely, we set $c_0 = \frac{1}{P(x_0)}$ and $c_i = \frac{P(x_0 \ldots x_{i-1})}{P(x_0 \ldots x_i)}$ for i=1, . . . , n-1 where n is the input length. As noted in section §4.2, the logarithm of c_i gives us the *surprisal* value for the input symbol x_i.

Rescaling factors are only introduced in the forward pass, during which the outer probability of a state $[l, r]$: $X \rightarrow \alpha \, . \, \beta$ has already been scaled by factors $c_0 \ldots c_{l-1} c_r \ldots c_{n-1}$.[6] More importantly, when computing expected counts, scaling factors in the outer and inner terms cancel out with those in the string probability in Eq. (1), implying that rule probability estimation is unaffected by rescaling.

5.1 Parsing Time on Dense Grammars

We compare in Table 1 the parsing time (on a 2.4GHz Xeon CPU) of our parser (Earleyx) and Levy's. The task is to compute surprisal values for a 22-word sentence over a *dense* grammar.[7] Given that our parser is now capable of performing scaling to avoid underflow, we avoid converting probabilities to logarithmic form, which yields a speedup of about 4 times compared to Levy's parser.

Parser	Time (s)
(Levy, 2008)	640
Earleyx + scaling	145

Table 1: **Parsing time** (dense grammars) – to compute surprisal values for a 22-word sentence using Levy's parser and ours (Earleyx).

5.2 Parsing Time on Sparse Grammars

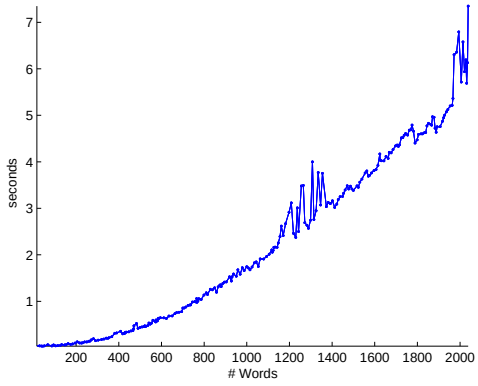

Figure 4: **Parsing time** (sparse grammars) – to compute Viterbi parses for sentences of increasing lengths.

Figure 4 shows the time taken (as a function of the input length) for Earleyx to compute a Viterbi parses over our *sparse* grammars (§3.2). The plot confirmed our analysis in that the special structure of our grammars yields approximately linear parsing time in the input length (see §3.3).

[5]Parser code is available at `http://nlp.stanford.edu/~lmthang/earleyx`.

[6]The outer probability of a state is essentially the product of inner probabilities covering all input symbols outside the span of that state. For grammars containing cyclic unit productions, we also need to multiply with terms from the unit-production relation matrix (Stolcke, 1995).

[7]MLE estimated from the English Penn Treebank.

6 Grammar Induction

We employ a Variational Bayes (VB) approach to perform grammatical inference instead of the standard Inside Outside (IO) algorithm, or equivalently the Expectation Maximization (EM) algorithm, for several reasons: (1) it has been shown to be less likely to cause over-fitting for PCFGs than EM (Kurihara and Sato, 2004) and (2) implementation-wise, VB is a straightforward extension from EM as they both share the same process of computing the expected counts (the IO part) and only differ at how rule probabilities are reestimated. At the same time, VB has also been demonstrated to do well on large datasets and is competitive with Gibbs samplers while having the fastest convergence time among these estimators (Gao and Johnson, 2008).

The rule reestimation in VB is carried as follows. Let α_r be the prior hyperparameter of a rule r in the rule set $\mathcal{R}$ and c_r be its expected count accumulated over the entire corpus after an IO iteration. The posterior hyperparameter for r is $\alpha_r^* = \alpha_r + c_r$. Let ψ be the digamma function, the rule parameter update formula is: $\theta_{r:X\to\lambda} = \exp\left[\psi\left(\alpha_r^*\right) - \psi\left(\sum_{r':X\to\lambda'} \alpha_{r'}^*\right)\right]$.

Whereas IO minimizes the negative log-likelihood of the observed data (sentences), -$\log p(\boldsymbol{x})$, VB minimizes a quantity called *free energy*, which we will use later to monitor convergence. Here $\boldsymbol{x}$ denotes the observed data and $\boldsymbol{\theta}$ represents the model parameters (PCFG rule probabilities). Following (Kurihara and Sato, 2006), we compute the free energy as:

$$\mathcal{F}(\boldsymbol{x},\boldsymbol{\theta}) = -\log p(\boldsymbol{x}) + \sum_{X\in\mathcal{N}} \log \frac{\Gamma\left(\sum_{r:X\to\lambda} \alpha_r^*\right)}{\Gamma\left(\sum_{r:X\to\lambda} \alpha_r\right)} - \sum_{r\in\mathcal{R}} \left(\log \frac{\Gamma\left(\alpha_r^*\right)}{\Gamma\left(\alpha_r\right)} + c_r \log\theta_r\right)$$

where Γ denotes the gamma function.

6.1 Sparse Dirichlet Priors

In our application, since each topic should only be associated with a few words rather than the entire vocabulary, we impose sparse Dirichlet priors over the $\texttt{Word}_t$ distributions by setting a symmetric prior $\alpha<1$ for all rules $\texttt{Word}_t\to w$ $(\forall t \in T, w \in W)$, where W is the set of all words in the corpus. This biases the model to select only a few rules per non-terminal $\texttt{Word}_t$.[8] For all other rules, a uniform hyperparameter value of 1 is used. We initialized rule probabilities with uniform distributions plus random noise. It is worthwhile to mention that sparse Dirichlet priors were proposed in Johnson (2010)'s work that learns Latent Dirchlet Allocation topic models using Bayesian inference for PCFGs.

7 Experiments

Our experiments apply sentence- and discourse-level models to the annotated corpus of child-directed speech described in Section 3. Each model is evaluated on (a) *topic accuracy*—how many utterances are labeled with correct topics (including the null), (b) *topic metrics* (f-scores/precision/recall)—how well the model predicts non-null topical utterances, (c) *word metrics*—how well the model predicts topical words,[9] and (d) *lexicon metrics*—how well word types are assigned to the topic that they attach to most frequently. For example, in Figure 1, the model assigns topic `pig` to the entire utterance. At the word level, it labels *piggie* with topic `pig` and assigns null topic to *wheres* and *the*. See (Johnson et al., 2012) for more details of these metrics.

In Section 7.1, we examine baseline models that do not make use of social cues (mother and child's eye-gaze and hand position) to discover the topic; these baselines are contrasted with a range of social cues (§7.2 and §7.3). In Section 7.4, we evaluate the discourse structures discovered by our models.

7.1 Baseline Models (No Social Cues)

To create baselines for later experiments, we evaluate our models without social information. We compare sentence-level models using three different inference procedures—Markov Chain Monte Carlo (MCMC) (Johnson et al., 2012), Expectation Maximization (EM), and Variational Bayes (VB)[10]—as well as the discourse-level model described above.

[8] It is important to not sparsify the $\texttt{Word}_{None}$ distribution since $\texttt{Word}_{None}$ could expand into many non-topical words.

[9] Topics assigned by the model are compared with those given by the gold dictionary provided by (Johnson et al., 2012).

[10] To determine the best sparsity hyperparameter α for lexical rules (§6.1), we performed a line search over $\{1, 0.1, 0.01, 0.001, 0.0001\}$. As α decreases, performance improves, peaking at 0.001, the value used for all reported results

Model	Topic				Word			Lexicon			Energy
	Acc.	F_1	P	R	F_1	P	R	F_1	P	R	
MCMC	49.07	60.64	48.67	**80.43**	29.50	17.63	**90.31**	14.83	8.10	88.10	
VB	53.14	60.89	50.53	76.59	25.62	14.94	89.91	16.71	9.25	85.71	156719
discourse	51.02	59.40	48.60	76.35	23.86	13.82	87.33	15.05	8.27	83.33	150023
discourse+init	**55.78**	**60.91**	**52.15**	73.22	**29.75**	**17.91**	87.65	**21.11**	**11.95**	**90.48**	**149458**

Table 2: **Social-cue models.** Comparison of sentence- and discourse-level models (*init*: initialized from the VB sentence-level model) over full metrics. Free energies are shown to compare VB-based models.

Model	Acc.	Topic F_1	Word F_1	Lexicon F_1
MCMC	33.95	40.44	**20.07**	10.37
EM	32.08	39.76	13.31	6.09
VB	39.64	39.22	17.40	12.27
discourse	**40.63**	**42.01**	19.31	**12.72**

Table 3: **Baseline (non-social) models.** Comparison of sentence-level models (MCMC (Johnson et al., 2012), EM, VB) and the discourse-level model.

Results in Table 3 suggest that incorporating topic continuity through the discourse model boosts performance compared to sentence-level models. Within sentence-level models, EM is inferior to both MCMC and VB (in accordance with the consensus that EM is likely to overfit for PCFGs). Comparing VB and MCMC, VB is significantly better at topic accuracy but is worse at topic F_1. This result suggests that VB predicts that more utterances are non-topical compared with MCMC, perhaps explaining why MCMC has the highest word F_1. Nevertheless, unlike VB, the discourse model outperforms MCMC in all topic metrics, indicating that topic continuity helps in predicting both null and topical utterances.

The discourse model is also capable of capturing topical transitions. Examining one instance of a learned grammar reveals that the distribution under $\texttt{Discourse}_t$ is often dominated by a few major transitions. For example, `car` tends to have transitions into `car` (0.72) and `truck` (0.19); while `pig` prefers to transit into `pig` (0.69) and `dog` (0.24). These learned transitions nicely recover the structure of the task that caregivers were given: to play with toy pairs like `car/truck` and `pig/dog`.

7.2 Social-cue Models

We next explore how topic continuity interacts with social information via a set of simulations mirroring those in the previous section. Results are shown in Table 2. For the sentence-level models using social cues, VB now outperforms MCMC in topic accuracy and F_1, as well as lexicon evaluations, suggesting that VB is overall quite competitive with MCMC.[11]

Turning to the discourse models, social information and topic continuity both independently boost learning performance (as evidenced in Johnson et al. (2012) and in Section 7.1). Nevertheless, joint inference using both information sources (*discourse* row) resulted in a performance decrement. Rather than reflecting issues in the model itself, perhaps the increased complexity of the inference problem might have led to this performance decrement.

To test this explanation, we initialized our discourse-level model with the VB sentence-level model. Results are shown in the *discourse+init* row. With a sentence-level initialization, performance improved substantially, yielding the best results over most metrics. In addition, the discourse model with sentence-level initialization achieved lower free energy than the standard initialization discourse model. Both of these results support the hypothesis that initialization facilitated inference in the more complex discourse model. From a cognitive science perspective, this sort of result may point to the utility of beginning the task of discourse segmentation with some initial sentence-level expectations.

7.3 Effects of Individual Social Cues

The importance of particular social cues and their relationship to discourse continuity is an additional topic of interest from the cognitive science perspective (Frank et al., 2013). Returning to one of the questions that motivated this work, we can use

[11] Detailed breakdown of word f-scores reveals that MCMC is much better at precision, indicating that VB predicts more words as topical than MCMC. An explanation for such effect is that we use the same α for all lexical rules, which results in suboptimal sparsity levels for $\texttt{Word}_t$ distributions. For MCMC, Johnson et al. (2012) used the adaptor grammar software to learn the hyperparameters automatically from data.

	all	no.child.eyes	no.child.hands	no.mom.eyes	no.mom.hands	no.mom.point
MCMC	49.1/60.6/29.5/14.8	**38.4/46.6/21.5/11.1**	49.1/60.6/29.6/15.3	48.0/59.7/29.0/15.5	48.7/60.0/29.3/15.6	48.8/60.3/29.3/15.6
VB	53.1/60.9/25.62/16.71	**49.3/56.0/22.6/15.1**	52.9/60.4/26.2/16.2	51.5/59.1/24.6/16.3	51.9/59.2/25.3/16.3	52.9/60.6/25.5/16.6
discourse+init	55.8/60.9/29.8/21.1	**53.7/59.2/27.8/19.7**$^{*+}$	55.2/60.7/29.0/21.4^{+}	54.7/60.0/29.0/21.6	55.2/60.1/29.1/21.4	55.6/60.8/29.5/21.7

Table 4: **Social cue influence.** Ablation test results across models without discourse (*MCMC, VB*) and with discourse (*discourse+init*). We start with the full set of social cues and drop one at a time. Each cell contains results for metrics: topic accuracy/topic F_1/word F_1/lexicon F_1. For row *discourse+init*, we compare models with/without a social cue using chi-square tests and denote statistically significant results ($p < .05$) at the utterance (*) and word ($^+$) levels.

	none	child.eyes	child.hands	mom.eyes	mom.hands	mom.point
MCMC	34.0/40.4/20.1/10.4	**45.7/57.3/28.9/13.6**	34.0/40.1/20.1/9.7	33.8/40.2/19.9/9.7	35.6/42.8/19.8/10.0	30.6/35.5/18.1/9.2
VB	39.6/39.2/17.4/12.27	**47.2/53.0/21.9/13.9**	43.0/45.8/15.4/12.9	42.9/46.5/14.6/12.4	41.1/43.8/17.1/12.4	39.7/39.7/17.5/13.4
discourse	40.7/41.8/19.2/12.1	**47.8/55.4/22.8/14.2**$^{*+}$	44.6/50.8/20.3/13.1^{*+}	44.7/50.1/21.7/14.3^{*+}	42.7/46.4/19.0/11.6^{+}	38.7/40.2/16.6/11.9^{*+}

Table 5: **Social cue influence.** Add-one test results across models without discourse (*MCMC, VB*) and with discourse (*discourse*). We start with no social information and add one cue at a time. Each cell contains results for metrics: topic accuracy/topic F_1/word F_1/lexicon F_1. For row *discourse*, we compare models with/without a social cue using chi-square tests and denote statistically significant results ($p < .05$) at the utterance (*) and word ($^+$) levels.

our discourse model to answer the question about the role that the `child.eyes` cue plays in child-directed discourses. Johnson et al. (2012) raised two hypotheses that could explain the importance of `child.eyes` as a social cue: (1) caregivers "follow in" on the child's gaze: they tend to talk about what the child is looking at (Baldwin, 1993), or (2) the `child.eyes` cue encodes the topic of the previous sentence, inadvertently giving a non-discourse model access to rudimentary discourse information.

To address this question, we conduct two tests: (1) *ablation* – eliminating each social cue in turn (e.g. `child.eyes`), and (2) *add-one*, using a single social cue per turn. Table 4 and 5 show corresponding results for models *without* discourse (the MCMC and VB sentence-level models) and *with* discourse (*discourse+init* for the ablation test and *discourse* for the add-one test). We observe similar trends to Johnson et al. (2012): the childs gaze is the most important cue. Removing it from the full model with all social cues or adding it to the base model with no cues both result in the largest performance change; in both cases this change is statistically reliable.[12] The large performance differences for `child.eyes` are consistent with the hypothesis that caregivers are *following in*, or discussing the object that children are interested in – even controlling for the continuity of discourse, a confound in previous analyses. In other words, the importance of `child.eyes` in the discourse model suggests that this cue encodes useful information in addition to the intersentential discourse topic.

[12]It is somewhat surprising when `child.eye` has much less influence on VB than on MCMC in the ablation test. Though results in the add-one test reveal that VB generalizes much better than MCMC when presented with a single social cue, it remains interesting to find out internally what causes the difference.

7.4 Discourse Structure Evaluation

While the discourse model performs well using metrics from previous work, these metrics do not fully reflect an important strength of the model: its ability to capture inter-utterance structure. For example, consider the sequence of utterances in Table 6. Our previous evaluation is based on the *raw* annotation, which labels as topical only utterances containing topical words or pronouns referring to an object. As a result, classifying "there" as `car` is incorrect. From the perspective of a human listener, however, "there" is part of a broader discourse *about* the `car`, and labeling it with the same topic captures the fact that it encodes useful information for learners. To differentiate these cases, Frank and Rohde (under review) added a new set of annotations (to the dataset used in Section 7) based on the discourse structure perceived by human, similar to column *discourse*, .

Raw	Discourse	Utterance
car	car	come here lets find the car
	car	**there**
car	car	is that a car
car	car	the car goes vroom vroom vroom

Table 6: **Topic annotation examples.** *raw* (previous metrics) and *discourse* (new metrics).

We utilize these new annotations to judge topics predicted by our discourse model and adopt previous metrics for discourse segmentation evaluation: *a=b*, a simple proportion equivalence of discourse assignments; p_k, a window method (Beeferman et al., 1999) to measure the probability of two random utterances correctly classified as being in the same discourse; and *WindowDiff* (Pevzner and Hearst, 2002), an improved version of p_k which gives "partial credit" to boundaries close to the correct ones.

Results in Table 7 demonstrate that our model is in better agreement with human annotation (*model-human*) than the raw annotation (*raw-human*) across all metrics. As is visible from the limited change in the *a=b* metric, relatively few topic assignments are altered; yet these alterations create much more coherent discourses that allow for far better segmentation performance under p_k and *WindowDiff*.

	raw-human	**model-human**
a=b	63.6	69.3
p_k	57.0	83.6
WindowDiff	36.2	61.2

Table 7: **Discourse evaluation**. Single annotator sample, comparison between topics assigned by the raw annotation, our discourse model, and a human coder.

To put an upper bound on possible discourse segmentation results, we further evaluated performance on a subset of 634 utterances for which multiple annotations were collected. Results in Table 8 demonstrate that our model predicts discourse topics (*m*-h_1, *m*-h_1) at a level quite close to the level of agreement between human annotators (column h_1-h_2).

	r-h$_1$	**r-h$_2$**	**m-h$_1$**	**m-h$_2$**	**h$_1$-h$_2$**
a=b	60.1	65.6	70.4	72.4	81.7
p_k	50.7	51.8	85.1	84.9	89.7
WindowDiff	29.0	30.1	60.1	66.9	72.7

Table 8: **Discourse evaluation**. Multiple annotator sample, comparison between raw annotations (*r*), our model (*m*), and two independent human coders (h_1, h_2).

8 Conclusion and Future Work

In this paper, we proposed a novel integration of existing techniques in parsing and grammar induction to offer a complete solution for simultaneously modeling grounded language at the sentence and discourse levels. Specifically, we used the Earley algorithm to exploit the special structure of our grammars to achieve approximately linear parsing time, introduced a rescaling approach to handle very long input strings, and utilized Variational Bayes for grammar induction to obtain better solutions than the Expectation Maximization algorithm.

By transforming a grounded language learning problem into a grammatical inference task, we used our parser to study how discourse structure could facilitate children's language acquisition. In addition, we investigate the interaction between discourse structure and social cues, both important and complementary sources of information in language learning (Baldwin, 1993; Frank et al., 2013). We also examined why individual children's gaze was an important predictor of reference in previous work (Johnson et al., 2012). Using ablation tests, we showed that information provided by the child's gaze is still valuable even in the presence of discourse continuity, supporting the hypothesis that parents "follow in" on the particular focus of children's attention (Tomasello and Farrar, 1986).

Lastly, we showed that our models can produce accurate discourse segmentations. Our system's output is considerably better than the raw topic annotations provided in the previous social cue corpus (Frank et al., 2013) and is in good agreement with discourse topics assigned by human annotators in Frank and Rohde (under review).

In conclusion, although previous work on grounded language learning has treated individual utterances as independent entities, we have shown that the ability to incorporate discourse information can be quite useful for such problems. Discourse continuity is an important source of information in children language acquisition and may be a valuable part of future grounded language learning systems.

Acknowledgements

We thank the TACL action editor, Mark Steedman, and the anonymous reviewers for their valuable feedback, as well as Chris Manning for helpful discussions. This research was supported under the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593).

References

Alfred V. Aho and Jeffery D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling; Volume 1: Parsing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Yoav Artzi and Luke S. Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

John Aycock and R. Nigel Horspool. 2002. Practical earley parsing. *The Computer Journal*, 45(6):620–630.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179 –190.

James K. Baker. 1979. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132.

Jason Baldridge and Alex Lascarides. 2005. Probabilistic head-driven parsing for discourse structure. In *CONLL*.

Dare A. Baldwin. 1993. Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20:395–418.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *HLT-EMNLP*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Benjamin Börschinger, Bevan K. Jones, and Mark Johnson. 2011. Reducing grounded learning tasks to grammatical inference. In *EMNLP*.

S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *ACL-IJCNLP*.

Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris 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*, 63(4).

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *ICML*.

David L. Chen, Joohyun Kim, and Raymond J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37:397–435.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *CoNLL*.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *ACL*.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Joshi Aravind, and Bonnie Webber. 2003. D-ltag system: Discourse parsing with a lexicalized tree adjoining grammar. *Journal of Logic, Language and Information*, 12:261–279.

Michael C. Frank and Hannah Rohde. under review. Markers of topical discourse in child-directed speech.

Michael C. Frank, Noah D. Goodman, and Josh B. Tenenbaum. 2008. A Bayesian framework for cross-situational word-learning. *Advances in Neural Information Processing Systems 20*.

Michael C. Frank, Joshua B. Tenenbaum, and Anne Fernald. 2013. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, 9(1):1–24.

Jianfeng Gao and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *EMNLP*.

Dan Goldwasser and Dan Roth. 2011. Learning from natural instructions. In *IJCAI*.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *ACL*.

Peter Gorniak and Deb Roy. 2007. Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2):197–231.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuk. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Jerry R. Hobbs. 1990. *Literature and Cognition*. CSLI Lecture Notes 21.

Liang Huang and Kenji Sagae. 2010. Dynamic programming for linear-time incremental parsing. In *ACL*.

Frederick Jelinek and John D. Lafferty. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323.

Mark Johnson, Katherine Demuth, and Michael Frank. 2012. Exploiting social information in grounded language learning via grammatical reduction. In *ACL*.

Mark Johnson. 2010. Pcfgs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *ACL*.

Joohyun Kim and Raymond J. Mooney. 2012. Unsupervised pcfg induction for grounded language learning with highly ambiguous supervision. In *EMNLP-CoNLL*.

Joohyun Kim and Raymond J. Mooney. 2013. Adapting discriminative reranking to grounded language learning. In *ACL*.

Alistair Knott and Ted Sanders. 1997. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175.

Kenichi Kurihara and Taisuke Sato. 2004. An application of the variational bayesian approach to probabilistic contextfree grammars. In *IJCNLP Workshop Beyond Shallow Analyses*.

Kenichi Kurihara and Taisuke Sato. 2006. Variational bayesian grammar induction for natural language. In *ICGI*.

Tom Kwiatkowski, Luke S. Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic ccg grammars from logical form with higher-order unification. In *EMNLP*.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Joop M. I. M. Leo. 1991. A general context-free parsing algorithm running in linear time on every lr(k) grammar without using lookahead. *Theoretical Computer Science*, 82(1):165–176.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *ACL-IJCNLP*, pages 91–99.

Percy Liang, Michael I. Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Association for Computational Linguistics (ACL)*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, FirstView:1–34.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *EMNLP*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. The rhetorical parsing of natural language texts. In *ACL*.

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *ACL*.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Livia Polanyi, Chris Culy, Martin Van Den Berg, Gian Lorenzo Thione, and David Ahn. 2004. A rule based approach to discourse parsing. In *SigDIAL*.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *EMNLP*.

Lawrence R. Rabiner. 1990. A tutorial on hidden Markov models and selected applications in speech recognition.

Remko Scha and Livia Polanyi. 1988. An augmented context free grammar for discourse. In *COLING*.

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

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *IJCAI*.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *NAACL*.

Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *NAACL*.

Stefanie Tellex, Thomas Kolla, Steven Dickerson, Matthew R. Walter, Ashis G. Banerjee, Seth Teller, and Nicholas. Roy. 2011a. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76.

Stefanie Tellex, Thomas Kolla, Steven Dickerson, Matthew R. Walter, Ashis G. Banerjee, Seth Teller, and Nicholas Roy. 2011b. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *AAAI*.

Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child development*, pages 1454–1463.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *ACL*.

Yuk Wah Wong and Raymond J. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *ACL*.

Chen Yu and Dana H. Ballard. 2004. On the integration of grounding language and learning objects. In *AAAI*.

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

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *UAI*.

Michael Zettlemoyer, Luke S.and Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *EMNLP-CoNLL*.