Learning Early Communicative Acts

Nancy Chang
CS 281, Spring 1998
Prof. Stuart Russell

1 Introduction

The acquisition of human language may be justly viewed as less than fully amenable to machine learning techniques. Although young children amply demonstrate the pervasiveness of the phenomenon, they serve rather less well as laboratory subjects: even if their input stimuli and output behaviors could be reliably and completely observed, their internal states and processes remain to a large extent opaque. The relative inaccessibility of data has given rise to theoretical accounts that clash in their most basic assumptions about the nature and nurture of the language capacity: What is the starting point for the acquisition process? To what aspects of their environments are children attuned during critical learning periods? What exactly is being learned, and how should this knowledge be represented? None of these has anything approaching a clear-cut answer yet, but the choices they present have major consequences for the learning process, as well as for longstanding controversies in linguistic theory surrounding the nature of grammar.

We can sidestep some of those controversies by focusing on the pragmatics of communication during the pre-linguistic and one-word stage. At these early stages, we need not tackle grammatical idiosyncracies and theoretical disputes head on, since the child cannot easily be viewed as acquiring abstract formal syntax. Rather, we can view pre-linguistic and early utterances as communicative acts that are performed in some environment, contingent on its particular goals and percepts, and that can result in specific effects. The word *up*, for example, appears quite early in English-speaking children as a request to be picked up, presumably on the basis of repeated exposures to the word in adult speech in association with just such an action. A child can thus be viewed as a learning agent with some complex representation of the world state and its own state, and words that become associated with those states (or changes to those states) form the basis for early communicative acts.

Recast this way, the problem is not one of inducing a specific formal grammar but rather one of learning behaviors that happen to involve linguistic stimuli and output. This paper shows how this functional approach to language acquisition facilitates the computational modeling of some data from the psychological and linguistic literature. More specifically, an on-line probabilistic learning algorithm called Bayesian model merging that has previously been applied to word learning is used to model data from the acquisition of early one-word utterances, especially Bloom's (1973) study of early function words. Results of this experiment and their implications for future modeling efforts are also discussed.

2 Early language acquisition as a machine learning problem

This section presents some facts from early child language and some of the constraints they suggest for a simple computational model of the acquisition of single-word communicative acts. Gross simplifications of complex phenomena have been made, but to the extent possible the assumptions are reasonable reflections of ideas in the literature.

2.1 Language acquisition from a pragmatic perspective

Since the focus here is on the one-word stage, a brief description of the development of communicative abilities up through that stage has been included as Appendix A. Observations of particular importance include the following:

- Language use begins toward the end of the sensorimotor stage, when children have gained control over their own motor and perceptual abilities. They have familiarity with and expectations about a variety of objects, settings and people.
- Children's early knowledge is structured by their participation in routine **events** controlled by caretakers and situated in a particular cultural, social and physical context (e.g., meals, baths, play). They know which participants play which roles in an event, with what kinds of results (usually the change of location or some other property of a central participant). It is these events (and not, for instance, objects) that are claimed to be primary in cognitive development.
- Children use gestures (verbal and otherwise) pragmatically, that is, to achieve their goals. This pragmatic knowledge includes the mastery of some basic principles of interaction (e.g., turn-taking) and communication (e.g., intonation, basic speech acts).

Thus, by the time they start using (recognizable) words, children have accumulated extensive and richly structured prior knowledge about events in the world and interactions with others. This pragmatic orientation is also evident in many of children's earliest words. Although many early words are object labels or person names, some of the most frequently occurring words do not have physically available referents and thus cannot be analyzed as purely ostensive. Rather, **function words** like *no*, *uh-oh* and *bye-bye* require notions of social interaction and the achievement (or failure) of goals. Appendix B describes function words in more detail and includes examples of their use in child and parent speech.

The hypothesis explored here is that appropriate representations of event-based prior knowledge, along with assumptions about how language is used, could render tractable the acquisition of early function words from examples. That is, children learn the meanings or use of function words from input examples consisting of an event representation and accompanying function word. The word *up*, for example, becomes associated with lifting (or change of location) events, and *hi* with arrivals. Such a scheme depends heavily on the precise nature of the input representation; the remainder of this section examines the machine learning issues in more detail.

2.2 Data "collection"

The most glaring problem for cognitive modeling tasks of the sort described here is the dearth of reliable and complete training data. Many diary studies have documented children's production and to some degree their comprehension, but these tend to focus on stages later than that of interest here (pre-linguistic and early one-word), and on what the child says rather than what the child hears. Such studies can thus be seen as snapshots of a particular stage of lexical development, more like testing data than training data.

Training data necessary for the learning task at hand – event experiences in context – have not yet been completely and systematically documented anywhere. This project thus relies on artificially constructed training examples, drawing heavily from Bloom's (1973) transcripts of her child between 16 and 22 months, which include occasional remarks about typical parental usage. These transcripts often (but not systematically) indicate values for even seemingly inaccessible features like goals, which are quite salient to a child but sometimes less so to the observer. See Appendix B for examples of transcript data.

2.3 Input representation and prior knowledge

Much of the difficulty in formulating language acquisition in computational terms is representational: infants inhabit a dynamic world with many continuous percepts, and how they represent and process the fluid sensations that make up their formative experiences, as well as the extensive prior knowledge they bring to the task, is as yet poorly understood. The observations made in Section 2.1 can be used to simplify the task in a cognitively motivated way. Using events as the basic unit of experience, I assume that children's perceptual inputs are pre-processed into a set of discrete attributes that describe (or parameterize) an event, its situational context and the accompanying (parental) speech act. That is, the space of possible events is featurized, and input data is translated from transcript form to a set of discrete features.

The feature set used here is listed in Table 1, along with a few examples. In general, these choices implicitly represent some of the prior knowledge discussed earlier. I assume, for example, that whole sets of

Attribute	Values	Examples			
		Context description			
activity	eat play read neutral	play	eat	neutral	neutral
location	ground sofa chair parent table absent present	sofa	table	absent	ground
desired activity	eat play read neutral	play	none	play	neutral
desired location	ground sofa chair parent table absent present	ground	none	none	parent
desired change	up down here absent none	down	none	none	up
		Event description			
schema	move appear disappear none	move	disappear	appear	move
agent	me parent	me	me	parent	me
patient	me parent food toy book	me	food	parent	me
patient animacy	true false	true	false	true	true
change	up down here absent none	down	away	here	up
achievement	true false na	true	na	na	true
		Utterance description			
speech act	comment query social	comment	query	social	comment
speech time	before during after	after	after	after	before
word	up down away down uh-oh bye-bye hi there more	down	gone	hi	up

Table 1: Attributes and possible values

complex stimuli might be summarized in one higher-level symbol, e.g. sofa, or patient animacy. In addition, some features (e.g., agent, patient, change) have meaning only in terms of a separate dynamic representation of events that they are assumed to parameterize. The exclusion of many perceptual inputs implicitly recognizes that the child already has some notion of what features may be relevant. These decisions have obvious simplifying consequences for computation and learning, but they also correspond to the conceptual prerequisites that have been hypothesized in the literature.

The context description includes both the current state of the world around them and their current goal(s).[1] Note that the **location** values are not mutually exclusive; some are more specific (ground) than others (present), and some are defined more statically (chair) than others (absent). This is because different event types (see below) have different sets of relevant location descriptions. The same is true for the goal-based **desired location** and **desired change**.[2] For the event description, I assume a small set of action or **schema** types, which are initiated by some **agent** and affect some **patient** (where the agent and patient may be the same, as in self-propelled motion). In general, agents are assumed to be people (the child or the parent), but patients may be other entities in the child's world, like food items or toys, and therefore may vary in **animacy**. Except for schema type none (corresponding to a static scene), events may also result in some **change** (of relative location or presence). This change may result in goal **achievement** (not applicable (na) when the child has no salient goal). The attributes **speech time** and **speech act** relate to the time (relative to action) and type of the utterance, respectively. I assume these features to be familiar to the child based on parental intonational cues and general pragmatic context.

I also make some simplifying assumptions about the linguistic input, which is in the form of a discrete symbol corresponding to a single **word**. This word is assumed to correspond to the stressed word of a longer utterance; stressed segments are more salient to the child (Foster 1990), and children as young as eight months old can successfully segment fluent speech into separate words on purely statistical grounds (Saffran *et al.* 1996). Also, all input is assumed to be child-directed, with little noise in the recognition process and joint attention with the parent established (Tomasello 1995). While real input is bound to be noisier, clear speech in focused parent-child interaction is likely to be more accessible to the child and thus have a greater impact on early word learning.

[1] This representation can be seen as an approximation of some of the many types of "context" distinguished by Nelson (1996): activity context, agenda context (goals), communicative act context and action context (event description) are directly represented, while social context (determined by location and people, which is limited here to one parent) and affective context (generally positive in the data recorded) are not.

[2] In retrospect it may have been better to make absence/presence a separate binary feature for both current location and desired location, and to distinguish clearly between location of child and location of other entities. See discussion in Section 4.

2.4 Feedback and hypothesis representation

Many formal approaches to language acquisition as grammatical induction, notably Gold (1967), were inspired by the longstanding observation that children are exposed to continual positive examples of language use but very little explicit negative feedback in the form of parental corrections. This task is similarly limited to positive examples, not of strings of words but rather of contextualized communicative acts as represented in the last section.[3]

The "vocabulary" or lexicon acquired from these examples is likewise contextualized: the hypothesis for a particular word is a generalization over the contexts in which it has been heard. In many cases, however, a single word may appear in multiple distinct (though possibly related) contexts, and these contexts may be diverse enough that a single generalization over all instances would not serve as a useful model of the word's use. Multiple word **senses** may be posited to account for such **polysemy**. All word senses can be represented in nearly the same form as the training examples (feature structure), where instead of particular values each attribute is associated with a distribution over possible values.

2.5 Testing

The hypothesis representation just described lends itself to more than one form of testing. If the associated word labels are considered classes into which particular events are grouped, the obvious test for standard feature-based classification (i.e., performance of the set on unseen examples) applies: Given the scene described in terms of the feature set, the system should emit the most probable associated word. This standard test has a clear correlate in children's actual early production patterns: children's tendency to comment on their own behavior or that of others (often to direct attention or inform others of their intentions) depends on classifying new events (possibly in progress) to find the most closely associated word.

A second mode of testing corresponds to comprehension; given contextual features (internal goals and external world state) and a particular word, the system should fill in the event description of the most compatible word sense. (The analogous behavior of the child is to either comply with a parental request or predict what is about to take place.) This corresponds to inference from incomplete data; since the known and unknown values loosely correspond to values available before and after an event, respectively, this task might also be considered a form of prediction.

2.6 Summary of learning task

- **The task:** Learn associations between words and the experiences with which they co-occur, in the form of probability distributions over feature-based experience descriptions. These learned associations are the basis for early communicative acts, since children can use them to perform more general actions like requesting and commenting.

- **The training experience:** Pairs of scenario descriptions (which include information about the context, event and utterance) and words, all discrete-valued.

- **The performance measure:** (a) Recognize a novel experience description with the most highly associated word; and (b) infer (predict) from initial context and word the most likely associated event.

The task as described involves both classification and inference (or prediction) from incompete data, so learned representations must be flexible enough to facilitate more than one mode of language use (comprehension and production) and more than one type of communicative act. Note also that (word) classes are not pre-specified but rather determined on the basis of data; potentially overlapping; and hierarchical, i.e., themselves divided into smaller subclasses.

[3]Note that the attributes used here include a different form of feedback: the feature **achievement** indicates whether the child achieved a goal during the communicative event or its associated action (if any). In general this kind of feedback may prove important enough to treat as an explicit reward, as discussed in Section 4.

3 Learning using Bayesian model-merging

The learning problem described is relatively simple from a computational perspective and could be handled in a variety of ways. Some additional constraints can be taken from the learning course of children: chilren's learning is on-line, incrementally incorporating new examples into existing representations; children exhibit **fast mapping**, or the ability to use or comprehend new words on the basis of very few examples; and production is conservative, often occurring only in familiar contexts and with few outright "mistakes" (i.e., uninterpretable to even parents, the most cooperative interlocuters).

The remainder of this paper explores the use of Bayesian model merging (Omohundro 1992), a cognitively motivated learning algorithm that satisfies the constraints of children's patterns of both acquisition and use. Although originally formulated in the context of curve approximations and radial basis functions, the model merging algorithm has since been extended to hidden Markov models and probabilistic attribute grammars (Stolcke 1994, Stolcke & Omohundro 1994). Most relevantly, Bailey (1997) applied model merging to the learning of action labels based on a set of features, including motor control parameters (force, posture, etc.) and world state (object size, object shape, etc.) that together describe an action. Since the bulk of this project is in the problem definition, and since the resulting problem is quite similar in some respects to action label learning, I adopt model merging in much the same form as described in detail in Bailey (1997) and its accompanying implentation.[4] I briefly describe Bayesian model merging and then provide results of some simple experiments.

3.1 Bayesian model merging

The model merging algorithm is based on the idea that initial learning is heavily dependent on remembering specfic training examples, but later training and performance might benefit from more general models that compactly encode a wider range of data. In the current context, the "model" is a set of features corresponding to a word (which here describe an event); the set of learned models (with probability distributions on each feature) is the lexicon. The algorithm creates a new model for each example encountered and merges similar models to improve the current lexicon. Lexicon improvements are judged according to the Bayesian criterion of maximum posterior probability $P(L \mid T)$, where L is the lexicon—or collection of word senses—and T is the training set. Applying Bayes' rule, this is equivalent to $P(L)P(T \mid L)$, or the product of the (prior) probability of the lexicon $P(L)$ (using a Dirichlet prior over multinomial feature values, and a structural prior favoring a more compact lexicon) and the likelihood $P(T \mid L)$ that a particular lexicon accounts for the training examples observed. While the optimal merge would be the one that most improves the posterior, the current implementation chooses merges based on the simpler but more efficient metric of similarity, determined using a modified chi-square statistic. (See Appendix C for a summary of the algorithm.)

The result of a merge is a model whose probability distributions for each feature are generalizations of those of the corresponding feature in the merged models. Here, all distributions are discrete, represented by counts of samples on each possible value (some of which are virtual samples), so "generalizing" in this context simply means adding the counts of each of the merged models. Merging thus produces more compact lexicons (with more peaked associated prior probabilities) that have more general word senses (with lower associated likelihoods of generating the training set). The model merging algorithm attempts to find the optimum trade-off between compactness and specificity, stopping when any merge would lower the posterior probability. At any point in training, however, the model set performs reasonably well, especially on input similar to seen examples.

Note that the model merging algorithm satisfies the constraints set out so far – namely, it learns word sense models based on examples, where models represent bidirectional, probabilistic, many-to-many mappings between words and situation descriptions. The representations are thus more flexible and interpretable than associations formed by standard backpropagation neural networks, and more explicit and general than the "representations" formed by clustering approaches. Disadvantages of this choice may also be apparent; though model merging has proven successful for Bailey's action label learning task, it is not clear that the same mechanisms and assumptions are entirely appropriate for this domain. For example, the separate

[4]The original implementation was by David Bailey, with later modifications by Bailey, Oliver King and myself; for this project some changes were made to improve dataset generation and testing facilities.

merging of each feature and the separate merging for different words reflect some independence assumptions that may be less valid in this domain than in those previously attempted. The current project can thus be seen as an attempt to determine the limits of the analogy between the two tasks.

3.2 Experimental results

Experiments were conducted using 200 training examples spanning nine function words, generated as described earlier and divided into disjoint training (80%) and test (20%) sets. The recognition and inference tests described earlier are performed as follows: For recognition, a novel feature structure is labeled by finding the most similar of the current models and returning its associated word. For inference, input in the form of a word paired with an initial context description (that is, a subset of the entire feature structure) is compared with only the context portion of all word sense models for that word, and the word sense with the most similar context description is used to fill in the remainder of the features (i.e., the event description). Correctness in the latter case is rather subjective; the method chosen here is to feed the resulting experience description (the input context, and the output event) back into the recognition test and compare the resulting label with the actual label. This measure is not ideal, and is moreover somewhat biased toward correctness; I focus on recognition results here.

Learning curves generated for both recognition and inference on all test examples are shown in Figure 1. Recognition results are also given for a second testing mode that takes advantage of a parameter specifying a threshold value for the guessed word's posterior probability; no answer is emitted when that value is below the threshold, in loose analogy with a conservative young language user's silence. The "answered" percent correct counts only those examples for which the program offered an answer. (There was little difference between the two testing modes for the inference case.) Results are averaged across five runs with different random splits into training and testing data.

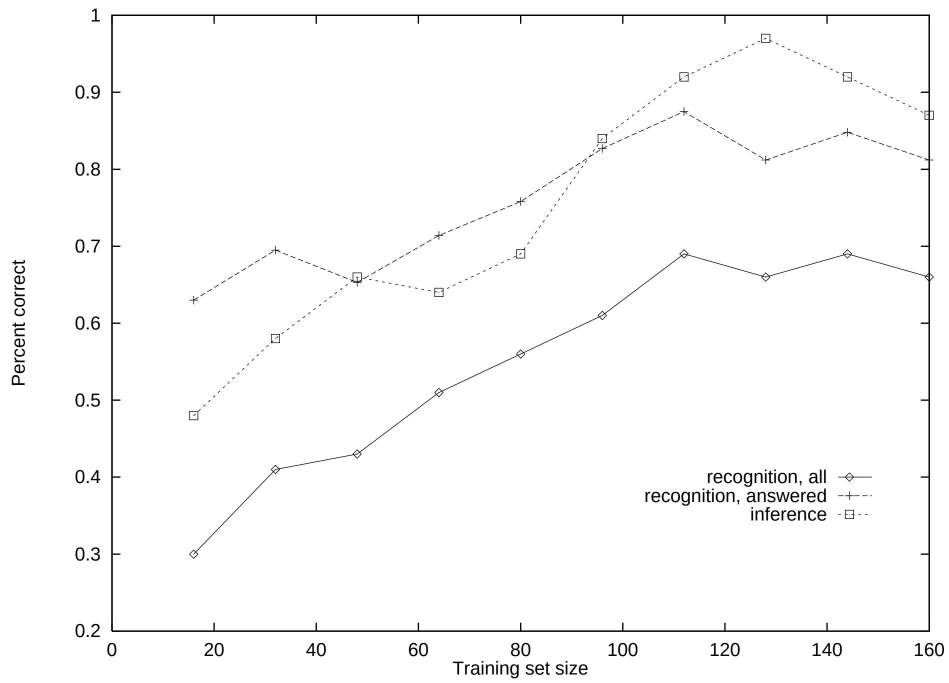

Figure 1: Performance on 40-example test set

Performance on both tests improved steadily with increasing examples, with evidence of overtraining consistently appearing after about 120 examples in the form of a decline or plateau in performance. Performance peaks at 69% for recognition (94% on only answered), and 97% for inference. All of these are far above chance performance (11.1%, since there are nine test words). This is also true even after only 20 examples (only 2 or 3 examples per word).

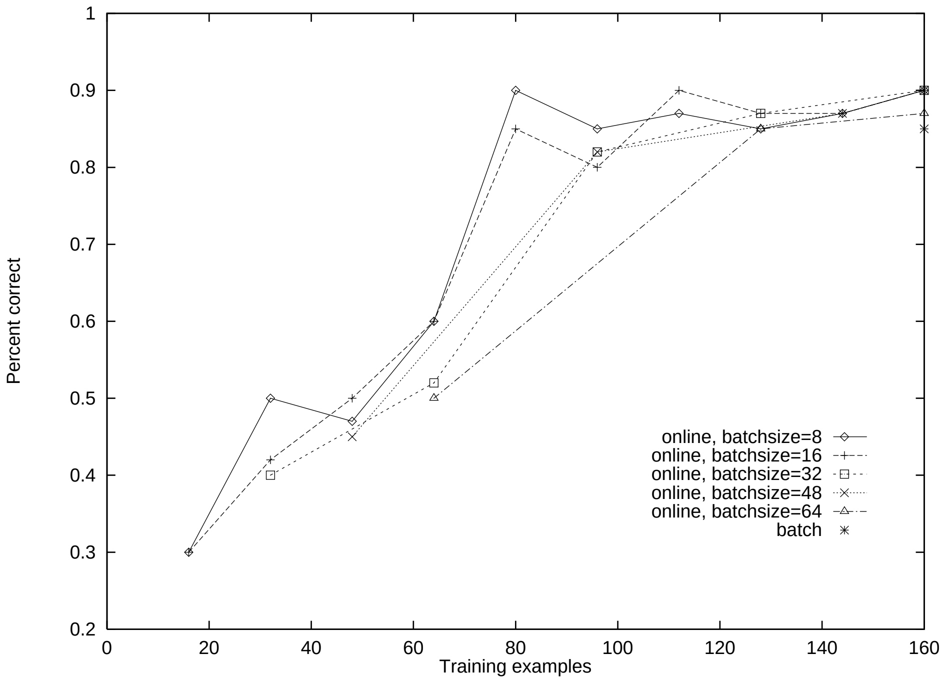

Figure 2: Batchsize experiments, recognition on all test examples

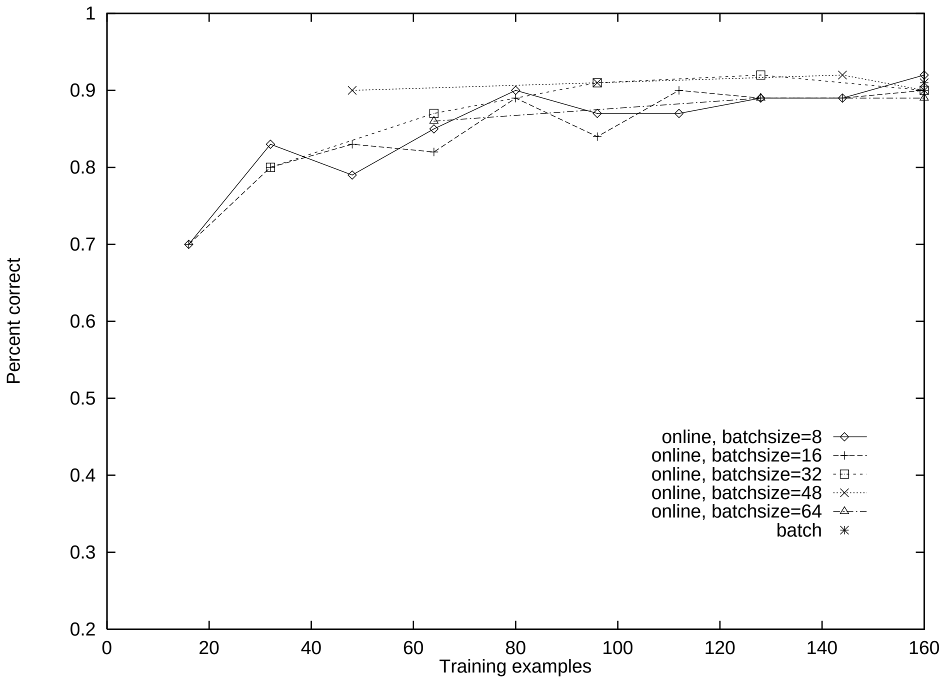

Figure 3: Batchsize experiments, recognition on answered test examples

A number of parameters for training and testing, such as minimum similarity required for a merge (not set) and number of virtual samples to create for each observed feature value in a training sample (typically 1), were set to default values for initial experiments. An additional set of experiments was conducted to investigate the effect of one of these, the **batchsize** parameter, which specifies how many new instances of a word to accumulate between merging episodes. Figure 2 shows results for the standard recognition test, and Figure 3 for the version counting only those answered. In general, smaller batchsizes seemed to yield slightly better recognition rates, and all online versions eventually outperformed learning using the entire 160-example batch on the standard recognition test. Overall, however, batchsize had no consistent pronounced effect, particularly in the answered recognition case. Interestingly, the largest leaps in performance seemed to occur for all batchsizes around 80-100 examples, or perhaps more consistently when the lexicon size had grown to around 14-18 senses.[5]

Learning curves are not themselves very illustrative for a task of this kind, however. A more qualitative judgment of the lexicon at the end of training shows that many of the acquired senses are intuitively reasonable, though not always what was expected. A lexicon with a total of 19 word senses learned in a typical training run is shown in Table 2. For words with multiple word senses, the sense covering the most example utterances is shown, with highest probability discrete value shown for each feature. Blank values indicate a feature deemed irrelevant for a particular word sense (based on a relatively flat distribution across values). Although only one word sense for each word is shown here, a few general characteristics of the lexicon are clear. First, learning produced only one or two word senses in most cases (seven of the nine words), with more senses generally corresponding to more examples in the input data. Second, a crude measure of feature relevance based on the peakedness of the corresponding probability distribution (blanks in the table indicate relatively flat distribution across values) reveals some patterns across words. In general, context description features were less consistently relevant than event description features, and the **schema** and **change** features are the only ones deemed relevant for every word sense.[6]

Attribute	**Values**								
				Context description					
word	away	bye-bye	down	gone	hi	more	there	uhoh	up
activity		play		eat		eat		play	
location	present	present	none	present	absent	absent	ground		ground
desired activity				none		eat	play		play
desired location		none		none				parent	sofa
desired change				none	none	here	up	up	up
				Event description					
schema	disappear	disappear	move	disappear	appear	appear	move	move	move
agent		parent	parent	parent	parent	parent	parent		me
patient		parent	me	food	parent	food	me		me
patient animacy	false	true	true	false	true	false	true	false	
change	away	away	down	away	here	here	up	down	up
achievement	true		true	na		true	true	false	true
				Utterance description					
speech act	comment	social	query	comment	social		comment	comment	comment
speech time	after	before	before	after	after	before	after	after	during
				Other					
number of examples	17	16	31	18	10	20	15	27	45
number of learned senses	2	2	2	2	1	3	1	2	4

Table 2: Summary of example learned lexicon.

The word *up* had the most learned senses as well as the largest representation in the training data; its senses are shown in Table 3. The uses of *up* observed in diary studies (see Appendix B) can be loosely correlated with the ones acquired here, as shown in italicized comments in the table. Both up1 and up2 are comments on some moving event resulting in an upward change of location and caused by either the child or the parent; the other, less common senses correspond to a parental query possibly preceding such an event (up3) or a comment on an elevated object (up4). Closer examination shows that the first two senses differ mostly in whether the child or the parent is the agent and thus might be better labeled "child-agentive" and "parent-agentive"; the terms "self-agentive" and "agentive" usually refer to whether the agent/causer of the

[5] Informal experiments showed learned lexicons to be similarly robust to variations of other parameters.

[6] The **speech time** feature is relevant for the senses shown in the table, but not for all senses in the lexicon, as shown in Table 3.

action is the same as the entity that is caused to move. Since that particular distinction would require a higher-order representation than the simple feature structure models used here, it is not too surprising that the acquired senses differ from what would be expected based on the linguistic or psychological evidence. Moreover, some values that might be expected to be irrelevant (and thus not set in the learned models) nevertheless appeared in the learned senses. Location and activity values in particular are not very salient in the adult's use of *up*. These values reveal some unexpected properties of the training data and may have disappeared with more data. The other words had simpler models. Overall, however, the various senses acquired do not diverge too much from those hypothesized by child language researchers.

sense	**up1**	**up2**	**up3**	**up4**
comment	*self-agentive*	*agentive*	*query*	*static*
number of examples	14	11	4	4
Attribute	**Values and probabilities**			
	Context description			
activity		read .533	neutral .375	neutral .500
location	ground .565		ground .230	high .384
desired activity	play .444	read .400	none .5	none .625
desired location	sofa .291	parent .333	none .357	none .357
desired change	up .681	up .526	none .416	none .416
	Event description			
schema	move .878	move .851	move=none .384	none .692
agent	me .937	parent .923	parent .833	
patient	me .421	me .562	me .555	toy .444
patient animacy		true .923	true .833	false .833
change	up .789	up .750	up=none .333	none .555
achievement	true .882	true .461		false .714
	Utterance description			
speech act	comment .833	comment .600	query .625	comment .625
speech time	during .588	before .642	before .714	

Table 3: Learned models for *up*. Blank spaces indicate distributions in which no value was peaked enough for the feature to be set and the feature is considered irrelevant.

An examination of the systerm's errors also reveals some interesting performance patterns. The errors made in both recognition and prediction are quite plausible: the system occasionally confused *up* and *down*, making an error that appears frequently in children's speech as well. The words *gone*, *away* and *bye-bye* are also often confused, since each has a usage involving the departure or disappearance of a person or item. This is a good illustration of the difficulty of evaluation: such errors should probably not be considered mistakes. Similarly, confusion between *there* and *up* is a natural product of dataset examples in which *there* is used to mark successful achievement (particularly after the child has climbed up into a new location). In all cases, the lexical entries learned by the system are either reasonable confusions of related concepts or not really mistakes at all but simply byproducts of models that overlap in semantic space; a more appropriate evaluation metric may take such facts into account.

4 Discussion

The experimental results described here indicate that simple learning techniques can be used to acquire reasonable approximations of the meanings of relatively complex words. More specifically, many-to-many mappings between function words and word sense models can be learned by children/systems with general conceptual abilities equivalent to those required for the assumed event representation. These conclusions are, however, subject to many caveats, as should already be apparent. In particular, strong assumptions have been made about the set of features available to children in the one-word stage. The learned word sense models are feature structures that provide a static, unstructured description of a dynamic, structured event. In other words, the active component of pragmatic knowledge has simply been featurized. Disregarding the obvious practical difficulties posed by extracting pragmatic features from transcripts or observation, this simplification might be a reasonable interpretation of evidence from studies of cognitive development. It also, however, allows us to skip a difficult step of the learning process: the association of changes in a particular state or world feature with particular words. As a result, although communicative acts were initially conceived as actions whose preconditions and effects must be learned, the model described here simply encodes **change** as a feature summarizing the relevant change of state.

As a result, the project is best viewed as an attempt to augment previous feature-based models of single-word learning to handle the pragmatic features needed for function words. Most directly, it builds on Bailey's (1997) strategy for learning verbs of hand actions by extending the parameters beyond ones that can be defined on biological grounds (force, direction, posture) to the more tenuous conceptual territory of goals, speech acts and context. By including these pragmatic features, the model is able to capture distinctions important for function words. The social content strongly associated with *bye-bye*, for example, plays an important role in distinguishing it from the otherwise conceptually similar *gone* and *away*; *there* and *uh-oh*, as markers of goal achievement and failure, respectively, are strongly associated with the aftermath of an event.

Note that these pragmatic features are simply additional conceptual features, not treated any differently from **animacy** or **schema**. While such an approach suffices for the current experiments, it is unrealistic in several respects. First, pragmatic information (especially the representation of goals and goal achievement) is likely to interact with the learning process in a different way from simple feature association. Children are more likely to learn when they care about an event and its results, where the amount they care may be represented as a reward (positive or negative), as in reinforcement learning. Training examples should thus be weighted by the presence and achievement of goals. Second, because the model does not distinguish between the propositional content associated with a word and its contribution to or use in a speech act, the model may acquire superfluous extra word senses corresponding to different speech acts. In the *up* example, for instance, up3 seems to correspond to examples in which an indifferent child is subject to the adult's query about picking the child up. It is not clear that this usage merits a different word sense. Moreover, some complications arise from the perspective switching in adult-child interaction: the current representation includes the perceived *parental* speech act, along with the *child's* goals. In general, dependencies between a speech act and its *speaker's* goals seem likely to matter, but capturing such dependencies would require more sophisticated representational machinery than those here, which are treated as independent and limited to propositions.

5 Conclusion and further work

At this early stage in the scientific understanding of language acquisition and use, the tools of machine learning seem most usefully devoted to bridging the gap between theory and data, both by providing at least a surrogate testing ground for hypotheses and by imposing a high standard of precision on both theoretical and empirical claims. Along these lines, this project explores the hypothesis that children's language-learning capacity is characterized in part by relatively simple learning mechanisms, combined with relatively more complex representational mechanisms and strong assumptions about prior knowledge and general cognitive/pragmatic abilities. Given that pragmatic notions have long posed some of the most unyielding challenges to linguistic analysis, it perhaps should not be surprising to find that much of the difficulty of this kind of project lies in the formulation of an appropriate computational framework that satisfies constraints imposed by the facts of early child language.

Using the representational framework specified here, a simple Bayesian model merging strategy did in fact acquire a reasonable lexicon of function words. The various simplifying assumptions made about the nature of input, however, make it unclear how well this approach would scale to other language learning tasks. Obtaining data poses an obvious problem: despite all attempts to base assumptions about data on psychological and linguistic evidence, in creating artificial data one inevitably toes the fine line between pre-processing the data and inadvertantly engineering the input representation toward expected results. This may not matter too much in the single-word case; featurizing complex events and relations seems to produce adequate results that reflect children's abilities passably well. Such an approach will most definitely not suffice for learning multi-word combinations. For that case it must be acknowledged that the linguistic portion of the input itself has rich structure that may include a combination of familiar and novel words, and that it is their combination must map onto the declarative structures of conceptual knowledge. A rather different learning problem emerges, since even the simplest multi-word constructions require that relations between lingustic units (e.g., word order) become associated with relations between corresponding portions of the event representation (e.g., agent or patient roles). Since these might be considered mappings between first-order relations, more sophisticated representational machinery and learning strategies will clearly be necessary.

Appendix

A Background: What kids know and when they know it

Even before a child is born, it can start to notice statistical correlations among its fledgling perceptual inputs, including the phonological and intonational patterns of the language spoken around its mother, as well as perceptual signals originating in its own body that are essential for the eventual development of motor control. The first several months of life inundate the child with all sorts of new sensory stimuli, some of which may be linguistic, and much of which is not, but all of which influence cognitive development.

Well before their first birthday, children begin to communicate through gestures, both vocal and non-vocal; these are summarized in Foster (1990). Reaching, for example, is used communicatively – that is, it functions to inform caretakers of the desire for some object – within as few as six months; raised arms to request to be picked up emerges around 10 months. Vocalized gestures (variously called proto-words, phonetically consistent forms or early vocables in the literature) perform the same function in a manner transitional between babbling and uttering recognizable adult words, including specific sounds that occur in consistent contexts and with consistent goals. Before long, prosodic features (rising/falling tones, simple stress) emerge, often corresponding to pragmatic elements of the communicative situation (e.g. whether a response is required or not).

Evidence of such knowledge – of the power of their own actions to effect the achievement of their goals, and of rudimentary conventions for interacting with others – also appears with abundance during the single-word stage, which usually begins around the nine-month mark and lasts for several months before the first word combinations appear (Bates 1976, Bloom 1973). In fact, it is often easier to categorize early words by their pragmatic context than by their underlying conceptual content (let alone grammatical category!); early one-word utterances by English-speaking children fall into a variety of adult grammatical classes (as do "up", "cookie" and "more"), but pragmatically each might be seen as a simple **request** (to be picked up, to be given something, etc.). The same words might in different pragmatic contexts function as **comments** on some aspect of the world around them, or perhaps specifically involve **social** interactions (as is clearest in the case of "hi" and "bye").[7]

B Function words and training data

Function words or **relational terms** are a subset of early utterances including words like *up, down, more, no, gone, bye-bye, away, there, uh-oh.* They have been found to appear consistently and frequently in the utterances of early English-speaking children. As many researchers have noted, these words cannot be interpreted as simple labels for static stimuli (as can nominals, like object labels and proper names). Rather, they tend to refer to some change of state, either realized or desired, suggesting that even at this early stage, the basic unit of cognition, and experience, is change- or event-based (Bloom 1973, Tomasello 1992, Nelson 1996). Bloom found in her 1973 study that children use these terms in context for a relatively limited set of conceptual notions. Children seem to notice – and talk about – location, movement, existence/presence,

[7] The discussion here centers on language production, which has for obvious reasons been easier to study than comprehension. Both production and comprehension data, however, provide only indirect evidence of the learning process; they can be viewed as snapshots of the child's stage of learning, where each utterance or interaction is one of an ongoing series of tests in which the child's overarching goal is simply to have its needs met.

change and recurrence. Further work has shown that function words often reflect the child's own needs and goals (Gopnik 1982) and a sensitivity to his or her immediate physical, social and activity contexts (Nelson 1996).

Function words appear common in parental speech, but usage at the stage in question is not well-documented. Some observations about parental and child usage based on comments from diary studies are shown in Table 4. Production data, including the contexts in which early function words are used, is much better documented, though in transcript form. Sample corpus data from Bloom's (1973) study are shown in Table 5. Note that these are records of usage during regular recording sessions and are only indirect indicators of the kinds of training evidence to which the child has been exposed.

word	parent usage context	child usage context
away	comment in eating/play *(Let's put it AWAY.)*	request; while looking for something
bye-bye	social situations; in play for disappearing objects	in play for moving or disappearing objects
down	putting child down	after sitting down; while trying to sit or descend; after object falls
gone	comment/question (often in eating context)	after finishing food; in reference to absent object
hi	social situations; in play for appearing objects	social/play
more	comment/question in eating context *(Here's MORE milk./Do you want MORE milk?/Let's read some more.)*	request for food recurrence; while reaching for food
there	comment on completion of activity	after completion of activity
uhoh	comment on sudden or startling events	after object or child falls
up	comment/question on child's movements	request to be picked up; while climbing up; reference to high object

Table 4: Function words and their typical usage patterns in parental and child speech

reference (page)	Mother/comment	Child
I1 (150)	(A trying to get up on chair)	up/
	Up. (A grunting)	
	(M helps A get on chair)	
	There.	
I63 (172)	(A tries to get on chair, then walks to other side of M; trying to make M get off the chair.)	down/ down/ up/
	Down?	
		up/ up/
	Up? Mommy off that chair?	
	(M stands up; A trying to get on chair where M was sitting; grunting)	down/
	(A gets on chair; sits)	up/ baby/
	Where's your baby?	
		gone/
	Gone?	
II24 (184)	([Toy] calf falls)	uh oh/
	What happened?	
	(Lamb and calf fall)	
	(A squealing, frustrated)	no/ no/
	Oh, what's the matter?	
	(A walks to M, trying to get in M's lap)	Mommy/
	Ah, honey, what's the matter?	
	(A trying to climb onto M's lap)	up/ up/
	You have a problem?	
		up/
	Up up.	
III48 (219)	(A drinks; puts down cup)	
	(A reaching for cup again)	more/
	(A looking into empty cup)	all gone/
	(A shakes cup; turns it over; looking at wet spot on floor)	uh oh/ again!/

Table 5: Examples of transcript data from Bloom (1973). 'A' stands for Allison, '/' for word boundary.

C Model merging algorithm

The following is a simplified on-line version of the model merging algorithm, following Bailey (1997):

incorporate example(model m, word v, lexicon L)
  create new word sense for m and add it to L
  when (k such senses have been created) **loop**:
    s_1, $s_2 \leftarrow$ best candidate merge [most similar pair of senses of v in L]
    old_post $\leftarrow$ compute L's posterior
    replace s_1 and s_2 in L with merge(s_1, s_2) [which sums counts on all feature values]
    new_post $\leftarrow$ compute L's posterior
    if (new_post $<$ old_post) **then** terminate loop
  endloop
end

The parameter k is the batch size, or the minimum number of examples to observe before merging begins, and is used to prevent merging too early.

References

BAILEY, DAVID R. 1997. *When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs.* Computer Science Division, EECS Department, University of California at Berkeley dissertation.

BATES, ELIZABETH. 1976. *Language and Context: The Acquisition of Pragmatics.* Academic Press.

BLOOM, LOIS. 1973. *One word at a time: the use of single word utterances before syntax.* The Hague: Mouton & Co.,.

FOSTER, SUSAN F. 1990. *The communicative competence of infants and young children: a modular approach.* London; New York: Longman.

GOLD, E.M. 1967. Language identification in the limit. *Information and Control* 16, 447–474.

GOPNIK, A. 1982. Words and plans: early language and the development of intelligent action. *Journal of Child Language* 9, 303–318.

NELSON, KATHARINE. 1996. *Language in Cognitive Development: The Emergence of the Mediated Mind.* Cambridge, U.K.: Cambridge University Press. Pp. 432.

OMOHUNDRO, STEPHEN. 1992. Best-first model merging for dynamic learning and recognition. Technical Report TR-92-004, International Computer Science Institute, Berkeley, CA.

SAFFRAN, JENNY R., RICHARD N. ASLIN, & ELISSA L. NEWPORT. 1996. Statistical learning by 8-month-old infants. *Science* 274, 1926–192.

STOLCKE, A., & S. OMOHUNDRO. 1994. Inducing probabilistic grammars by Bayesian model merging. In *Proceedings of Second International Colloqium on Grammatical Inference (ICGI-94)*, volume Lecture Notes in Artificial Intelligence 862, 106–118. Springer-Verlag.

STOLCKE, ANDREAS. 1994. *Bayesian Learning of Probabilistic Language Models.* University of California, Berkeley dissertation.

TOMASELLO, M. 1995. Joint attention as social cognition. In *Joint attention: Its origins and role in development*, ed. by C.D. Moore P. Hillsdale, NJ: Lawrence Erlbaum Associates. Educ/Psych BF720.A85.J65 1995.

TOMASELLO, MICHAEL. 1992. *First verbs: A case study of early grammatical development.* Cambridge, UK: Cambridge University Press.