Early Word Learning in Context

by Brandon C. Roy

Ph.D. Thesis Proposal, Media Arts and Sciences
Massachusetts Institute of Technology
October 20, 2011

Deb Roy
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Date: _____

Michael C. Frank
Assistant Professor of Psychology
Stanford University

Date: _____

Shimon Ullman
Professor of Computer Science
Weizmann Institute of Science

Date: _____

Contents

1 Introduction **2**

2 Background **3**

3 Thesis summary **4**
- 3.1 Activity contexts . . . 5
- 3.2 Activity contexts vs. Formats . . . 5
- 3.3 Could activity contexts shed light on language acquisition? . . . 6

4 Foundations **7**

5 Methods **8**
- 5.1 Methods for inferring hidden variables . . . 9
- 5.2 Other methods . . . 11

6 Evaluation and analysis **11**

7 Conclusion **13**

8 Research Plan **13**
- 8.1 Completed work . . . 13
- 8.2 Timeline . . . 14
- 8.3 Required resources . . . 14

9 Author Biography **14**

Early Word Learning in Context

Brandon C. Roy

October 20, 2011

Abstract

The remarkable achievements of early word learning and productive language use take place in the natural social setting of a child's first years of life. This thesis investigates early word learning in a naturalistic, dense, longitudinal corpus of one child's development over time, focusing on the contributions of environmental variables and social context. The key idea is that word learning is strongly supported by the social activity structures between child and caregivers. To explore the emergent structures and processes that support early word learning requires naturalistic, longitudinal data. In this thesis, the data consists of hundreds of thousands of hours of audio and video recordings. To work with this data, new methods for large scale speech transcription are developed. A detailed timeline of vocabulary growth is presented and correlated to environmental factors such as caregiver word usage. The proposed new work includes methods for extracting nonlinguistic "activity contexts" and linking these contexts to early word learning. An activity context is a variable that bundles together space, time, language and social participants to describe "what is happening" in daily life, and serves as an approximation to Jerome Bruner's (1983) idea of a *format*. Through the lens of activity contexts, questions such as which activities contribute more to word learning and how these activities develop over time can be addressed. By studying language acquisition in context, we hope to better illuminate the link between language, communication and the social world.

1 Introduction

Children are robust language learners, successfully learning across a wide variety of cultural and individual circumstances. This has led researchers to question the contributions of the child's innate faculties relative to the role of the environment. In the face of environmental variability, it is reasonable to suspect powerful innate mechanisms. Many laboratory studies have contributed to our understanding of these mechanisms, such as children's sensitivity to statistical regularities, object properties, and verbal and nonverbal cues. Yet to apply these mechanisms to learning, the child must be situated in a linguistic environment. The learning environment is not simply a pre-existing and immutable condition, but is influenced and even constructed by child and caregivers. How does the natural setting of a child's first years of life contribute to and support language learning?

This thesis begins to address these questions through analysis of a new dense, longitudinal, *naturalistic* corpus collected for the Human Speechome Project (D. Roy et al., 2006), which

captures one child's development from birth to age three. The basic perspective taken is that language, communication and context mutually support one another, and to study child language acquisition we must view it in context. The work of Jerome Bruner (1983) on the communicative support system jointly constructed and managed by child and caregivers serves as a guiding principle. The key element of this support system is the idea of a "format", which is a stable, predictable social structure through which child and caregiver can interact and communicate. For Bruner, formats served as a kind of "scaffold" for constructing language.

The bulk of this thesis consists of analyzing environmental factors in early word learning, beginning with investigations involving just the speech transcripts and correlating caregiver word usage patterns with the child's word learning. While this is an important part of the thesis, the present discussion focuses on new work to investigate language in context. In particular, this document emphasizes steps toward operationalizing the idea of formats and studying their role in word learning. To get at this idea, I propose to extract "activity contexts", which are semantically meaningful labels of *what* is happening in the child's experience. For example, `mealtime` is an activity that bundles together language use, location, time of day and other elements that may all contribute to constraining and simplifying the child's learning task. Rather than studying the child's exposure to a particular word in the sea of all language he has ever heard from birth, we can view that word with respect to salient contexts. There are numerous challenges to this proposal which are addressed in the following sections. But the overarching goal is to take advantage of the unique opportunity provided by the Human Speechome Project to study language acquisition in context and as it happens "in the wild."

2 Background

At the end of the 17th century John Locke wrote, in *An Essay Concerning Human Understanding*, "For if we will observe how children learn languages, we shall find that, to make them understand what the names of simple ideas or substances stand for, people ordinarily show them the thing whereof they would have them have the idea; and then repeat to them the name that stands for it; as white, sweet, milk, sugar, cat, dog." This picture presents the child as a kind of simple associator, learning the mappings from words to objects. However, there are many ambiguities even in "simple" association tasks, such as which object or property of the object is being referred to. For example, the words "white", "glass" and "milk" could all be used when referring to a glass of milk. This model of learning relies entirely on the adult to ensure the child is attending to the right object (and the right *aspect* of the object, action, or concept) when a word is used. In fact, children bring many biases to bear on these problems and a *core knowledge* of objects (Spelke, 1994; Spelke & Kinzler, 2007), as well as many skills, such as the ability to infer the referential intent of others when learning words. Baldwin (1991) tested whether children would associate a novel word to an object in plain view, or to a second object, attended to by the experimenter but hidden from the child's view. Contrary to what Locke may have predicted, children associated the word to the hidden object, supporting the idea that children can infer the referential intent of others.

The ability to infer the intent of others is a crucial social and communicative skill. By under-

standing other's intent we know what must be communicated and what is already understood. One way to understand the intention of others is by understanding their social role and the activity in which they are engaged. This is a key component of Bruner's (1983) "language acquisition support system" (LASS), made explicit in his idea of a *format*. A format is a stable structure in which both child and caregivers participate that guides the child toward understanding the "transactional" nature of language, social roles, and serves as a scaffold for language. It is "...a rule-bound microcosm in which the adult and child do things to and with each other." (Bruner, 1985). In Bruner (1983) the game of "peek-a-boo" is studied as a format between mother and child that begins before the child's first words and grows in sophistication as the child develops. The child eventually takes on the role as the primary actor, initiating the game and uttering the all important "peek-a-boo!" Over the course of development mother and child played this game, mutually understood and enjoyed by both participants, with the mother continually demanding more expressive language and behavior from the child. In the constrained world of peek-a-boo, words had particular meanings and effects on the state of the game and the other participant.

For Bruner, studying language acquisition in a controlled laboratory setting was fundamentally limited since it could not adequately capture the real conditions of language learning, namely, the LASS. Bruner (1983) says, "The issues of context sensitivity and the format of the mother-child interaction had already led me to desert the handsomely equipped but contrived video laboratory... in favor of the clutter of life at home." The "clutter of life at home" was just the setting for Esther Dromi's (1987) longitudinal study of her daughter. In this work, she makes a strong claim for the value of naturalistic, longitudinal data. Her detailed recording and analysis of her daughter's early lexical development revealed striking patterns of vocabulary growth, and provided evidence for a stage-like transition from single words to syntactically productive multiword speech.

McCall (1977) recognized both the challenges and the value of naturalistic data analysis to developmental psychology. He frames the problem around the question of "can versus does"; controlled experiments help answer the question "can X, under certain circumstances, lead to Y?" But such experiments may not necessarily answer "Does X, under normal circumstances, actually lead to Y?" It falls to naturalistic studies to address this important question. McCall laments that such studies are rare in developmental psychology, suggesting that this may be due to lack of experience with methods of analysis, the perception that they are "hopelessly confounded," or being overly costly and time consuming. However, work such as Bruner's and Dromi's have contributed methods for analysis and helped to demonstrate the value of this kind of data. The work here owes much to this tradition.

3 Thesis summary

The foundation of this work is the data collected and annotated for the Human Speechome Project. This corpus, described in more detail in the next section, consists of a dense, naturalistic, longitudinal audio and video record of one child's development.

The work in this thesis follows a trajectory beginning at raw audio and video and leading to an analysis of high-level contextual factors and their contribution to word learning. Building up

from this starting point, our methodology for annotation, specifically for speech transcription, will be presented. The first analysis work on vocabulary growth will be described, which involves both automatic, noise robust methods for identifying the child's first use of a word and manual annotation tools to ensure correctness. With a reliable vocabulary growth timeline it is then possible to link environmental factors to word learning. The earliest work on environmental factors looked at word frequency in caregiver speech and its relation to the child's age of acquisition (AoA), later work included other factors. These studies serve as an important link in the chain from raw data to rich contextual analysis, but they also provide baselines for evaluating and comparing the contributions of nonlinguistic context to word learning. The target point for this trajectory is an analysis of the high-level contextual factors and their contribution to word learning.

3.1 Activity contexts

"Contextual factors" and "nonlinguistic context" can refer to many things – here, I wish to focus on a restricted sense that loosely identifies what activities the child and caregivers are engaged in over short time ranges. I use the term $activity\ context$ to try to capture "what is happening" at the temporal granularity of minutes. Mealtime, story time, and playing with toys are possible activity contexts. Activity contexts are an attempt to capture some (though not all) of the salient elements of Bruner's formats, such as who is involved, $when$ and $where$ does it take place and potentially some aspects of the dynamics of speaker turn taking. What they do not necessarily capture is the deep structure of the format itself.

3.2 Activity contexts vs. Formats

Bruner's inquiry focused on the role of the "language acquisition support system" that emerges in natural child-caregiver interaction, and its key structure, the format. Formats have a consistent deep structure over time, while activity contexts are essentially clusters of audio and video that represent consistent activities or behaviors. A person viewing the audio and video might give all these episodes the same label, such as `breakfast` or `book-reading`. Assuming activity contexts can be extracted, what value do they have in understanding language in context, and how do they relate to Bruner's formats?

First, they provide a way of quantifying how the child spends his time. Building on McCall's argument, naturalistic data offers the opportunity to understand what $does$ happen in the course of daily life. How much time does the child spend looking out the window and naming cars as compared to book reading? Does this balance change over time?

Second, these contexts can be used as a lens into language by partitioning the corpus into more semantically related episodes. Language used during mealtime may have certain consistent elements that are lost when mixed with all language in the corpus. Prior work has shown the important relationship between a word's overall frequency in caregiver speech and the AoA of that word, but underlying this is a more subtle and interesting connection: when words are grouped by syntactic class, nouns are less frequent overall yet they are learned earlier and their

acquisition is more strongly correlated with caregiver usage frequency. The point here is simply that partitioning language use appropriately can reveal patterns that are otherwise obscured.

The third point is a practical one: to study the deep structure of a format and its attendant language use, one needs to extract examples of that format from the data. In Bruner's (1983) study, the primary focus was on the game of "peek-a-boo", and few other formats are discussed. The peek-a-boo format is fairly well defined by a single keyword, and occurrences should be relatively easy to find by searching the transcripts. But if the format idea is to be taken seriously then there must be other formats that are not as clearly defined by a keyword. So there are two problems: given a format, how to find examples in the data, and how to identify other formats?

3.3 Could activity contexts shed light on language acquisition?

Previous work investigating the contribution of input factors to early word learning found that $recurrence$ is one of the best predictor variables for age of acquisition (Vosoughi et al., 2010). Recurrence measures the frequency of word use within a short, roughly one minute period of time. D. K. Roy & Pentland (2002) found that mother's speech naturally contained frequent repetitions of salient words within close temporal proximity. The extent of such "redundant" speech was carefully studied in (Snow, 1972).

Recurrence might reflect caregiver's conscious or unconscious attempt to highlight particular words through repetition. It may also arise as a byproduct of language use during a focused activity. In (Vosoughi et al., 2010) no distinction was made between the possible sources of recurrence, but it seems reasonable to suspect that words with high recurrence values may be learned earlier not because of recurrence $per\ se$, but because they are used during focused activities. This is one clue that identifying language use with activity contexts may be a fruitful pursuit.

The work by Miller (2011) and Shaw (2011) has a more direct bearing on the question of how activity contexts might relate to language acquisition. In Shaw (2011), person tracking algorithms were developed and applied to the Speechome video data, linking speech transcripts to the locations where they were uttered. Measures of spatial "clusteredness" for words correlated with the child's age of acquisition. In Miller (2011) a more abstract representation of motion in video was developed, but it also enabled a link between speech transcripts and space. The key finding was that words whose spatial usage patterns differed from overall language use tended to be learned earlier. For a given word, the spatial distribution of its use could also be displayed, showing that certain words were tied to space in unique and meaningful ways. Figure 1 illustrates this for the word "sea" which was not only highly spatially localized, but learned early relative to what other non-spatial variables predicted.

The spatial location of word use may in itself be the crucial factor contributing to word learning, or, as with recurrence, it may be a reflection of a deeper underlying factor. The view taken here is the latter; that the crucial underlying factors are the human activities that would be better captured through activity contexts. Activity contexts may be strongly tied to locations, such as book reading in a particular chair, but they may also be tied to times of day, particular caregivers and so on.

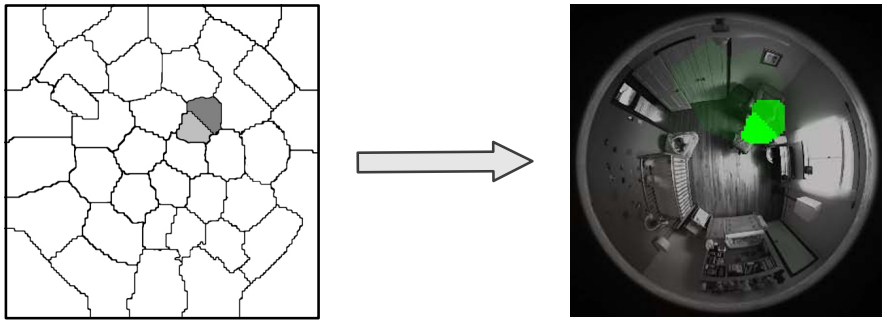


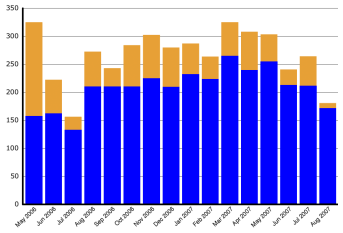
Figure 1: The relative spatial distribution of the word "sea". The picture on the left shows the region boundaries in the baby's bedroom. The gray regions have a high value for $\frac{\Pr(X|w)}{\Pr(X)}$, which measures the spatial activity distribution when a word is uttered, $\Pr(X|w)$, relative to the overall spatial activity distribution $\Pr(X)$. The picture on the right shows the highlighted regions superimposed on the video of the child's bedroom, which correspond to a chair where books are read to the child. Adapted from (Miller, 2011).

4 Foundations

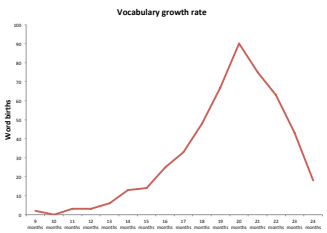
The foundation of this thesis is the audio and video recorded for the Human Speechome Project. The dataset consists of dense, naturalistic, longitudinal audio and video recordings of one child's life at home, from birth to age three. It is, by orders of magnitude, the largest dataset to date of a single child's development, consisting of more than 120,000 hours of audio and 90,000 hours of video. The entire home of the family was recorded from 11 cameras and 14 microphones for an average of 10 hours per day, and rather than selectively turning the system on to record, it was selectively turned *off* to maintain privacy. Thus, the recordings are unstructured, and much work has gone into annotating the data.

The key annotation required for studying language are speech transcripts. Toward this end new semi-automatic software was built for large scale, rapid speech transcription (Roy & Roy, 2009). To streamline transcription, speaker ID is instead performed separately, using a fully automatic system. The end result is that we have transcribed a significant portion (~80%) of the child's 9-24 month age range, which typically captures the emergence of first words up to multiword speech (Dromi, 1987). All analysis is based on the 9-24 month subset of the data. Figure 2(a) shows the amount of audio recorded and the amount transcribed per month.

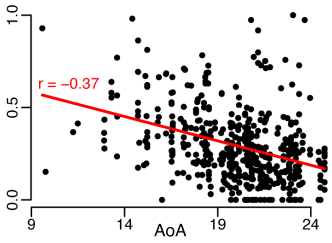
The child's vocabulary growth and relation to caregiver input speech was first studied in (Roy et al., 2009), and again with respect to caregiver prosody in (Vosoughi et al., 2010). The timeline of when each word entered the child's productive vocabulary, which we called a "word birth", is central to our study of word learning, and with significantly more transcripts now available the vocabulary has been re-derived using better, noise robust techniques. Noise robustness is crucial since automatic speaker ID is imperfect and transcripts can contain errors. In the case of the child's vocabulary, human annotators confirmed and corrected vocabulary growth timelines. Analysis of the child's linguistic environment revealed both new and expected



(a) Transcription progress for 9-24 months.



(b) Vocabulary growth rate



(c) AoA vs. recurrence

Figure 2:

factors contributing to word learning. Figures 2(b) and 2(c) show the child's vocabulary growth rate and the relationship of AoA to recurrence.

In general, the size of the Speechome Corpus has shifted the balance toward methods that can be applied at scale. Part of what makes this data so rich are the extensive video recordings, which hold the key to investigations of physical and social context. Two basic, automatic video annotations have been used extensively; person tracks and spatial activity distributions. Shaw (2011) developed a system to track people as they move within the camera view. Although tracks do not capture person identity and are error prone due to the difficulty of the task, in aggregate the tracks capture important movement patterns. Miller (2011) developed spatial activity distributions, which discretize the video into regions and capture the amount of movement in each region at a given time, so they are robust but imprecise. However, they do capture meaningful patterns of location and movement. In addition to the two automatic video annotations, we also manually annotate which video camera recorded the child over time, and whether he was awake or asleep. This is to ensure we transcribe "child-available" speech, but these annotations can be used for other purposes as well.

In sum, the raw audio and video, speech transcripts, speaker ID, video-channel annotations, person tracks, and spatial activity distributions make up the Speechome Corpus. In addition, the child's vocabulary growth timeline and some limited caregiver prosody data is available. Figure 3 presents an overview of the raw data, annotations and dependency relationships. While speech transcription is an ongoing process, existing annotations are incrementally refined, and automatic methods are occasionally improved, the current Speechome Corpus is a usable, unique and powerful resource for studying language acquisition as it naturally occurs.

5 Methods

Using activity contexts as an organizational principle for studying language requires identifying useful contexts and applying them to data. The scale of the Speechome Corpus makes this task especially challenging – manual techniques may be limited with thousands of hours of recordings, so automatic methods must be considered. However, the scale may also facilitate automatic

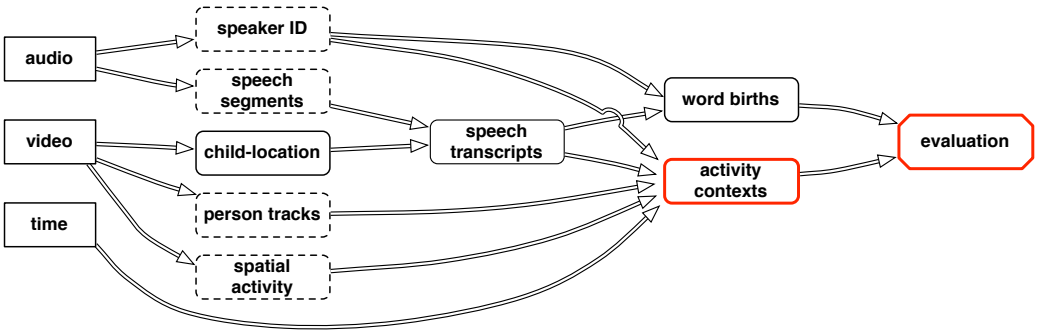


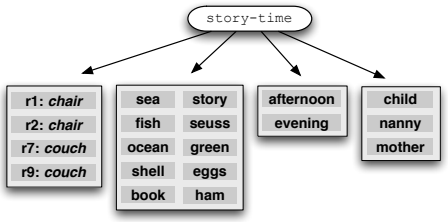
Figure 3: Raw data, metadata and analysis dependencies. Dashed boxes indicate that the data is acquired in a fully automatic fashion, solid boxes indicate that semi-automatic or fully manual methods are used. Audio, video and time are directly recorded raw data and have no dependencies. Annotations such as person tracks depend only on video, while speech transcripts depend on both speech segments and the child's location. Most of the new work described in this proposal is on activity contexts and associated evaluations, highlighted in red.

techniques since stable patterns should generally be well represented and not suffer from data sparsity.

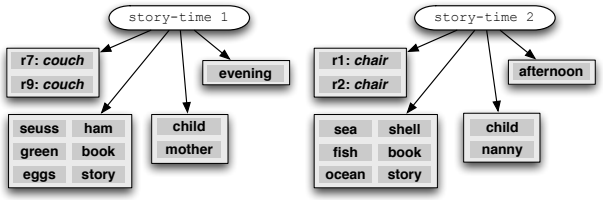
A path toward automatic discovery and labeling of activity contexts is suggested by the work of Miller (2011) and Shaw (2011), discussed in the previous section. The two key data sources they used were language (from transcripts) and space (from person tracks or video activity patterns.) Two additional data sources are speaker identity and time of day. These four dimensions: transcripts, location, speaker identity and time of day are effectively the "observables" in a dynamic process that unfolds over time. What binds these four multimodal, temporally aligned data streams together? The key idea is that these data streams are not independent and are instead reflections of a "hidden variable" that underlies and generates the observable data, namely, an activity context variable. Figure 4 provides an example for the activity context `storytime`. Stories are typically read to the child by his nanny in the afternoon and his mother in the evening. Afternoon stories are usually read in his bedroom chair before a nap, while evening stories are often read on the couch in the living room. Words such as "book" and "story" are common, as well as story specific words such as "sea", "fish", "green", "eggs" and so on. Figure 4 shows `storytime` as a single activity context, or possibly split into two, more specific contexts to model the distinction between afternoon and evening stories.

5.1 Methods for inferring hidden variables

A standard technique for learning with incomplete data is *Expectation Maximization* (EM) (Dempster et al., 1977). In this case, the idea is essentially to postulate a set of unknown activity variables and begin by assuming a random, hidden labeling of temporal regions with these variables. The learning process works by finding the maximum likelihood parameters for the activity variables given the observed data, and then updating the hidden labels given the new parameters in an iterative fashion.



(a) A representation of `storytime`, showing conditional distributions over locations, words, times and people.



(b) More specific representations of `storytime`, capturing stories read by the nanny in the afternoon as distinct from stories read by the mother in the evening.

Figure 4: Representations of the activity context `storytime`.

Hidden Markov Models (HMMs) are a specific example of how to model sequential data whose dynamics depend on hidden state variables. Rabiner (1989) describes HMMs for speech recognition but gives a good, general description of the learning and inference algorithms. Once an HMM has been trained, it can be used to infer the most likely sequence of hidden variables to have generated the observed data. There are many details that must be addressed if HMM techniques are to be applied, but they are a reasonable class of methods to consider.

While the observable data is inherently temporal, we need not explicitly model temporal dynamics to capture the idea of activity contexts. For this kind of modeling, we can look to techniques such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and more recently (and perhaps more appropriately) Latent Dirichlet Allocation (LDA) (Blei et al., 2003). These methods are typically applied to document modeling, where a document is viewed as a mixture of latent (hidden) topics, and the latent topics induce distributions over unordered sets of words. These methods are useful for modeling document collections because the learned hidden topics tend to collect semantically related groups of words, and documents tend to be represented as sparse combinations of topics. The most straightforward extension from document collections to the Speechome Corpus is to view contiguous blocks of time as "documents." From this perspective, the "words" are the events that occur within the time block, which includes actual words from transcripts, the speakers, the time of day and location. A potential problem with this extension is that standard LDA is designed for discrete data, but some of the activity context data modalities may not be well suited to discrete representations.

Prior work in our research group used related techniques for unsupervised indexing of sports video (Fleischman & Roy, 2007). In this work, video features from televised baseball games were extracted and linked to the closed-caption text using a model based on the author-topic model (Steyvers et al., 2004). The goal in this case was indexing of video content, but the approach is relevant to discovering activity contexts. Latent variable models applied to multimodal data, in this case text and images, is described in (Barnard et al., 2003) and (Blei & Jordan, 2003). In (Purver et al., 2006), topic modeling and segmentation of temporal data are treated together, and the authors' topic segmentation model compares favorable to an HMM.

While there are successful precedents for learning latent variable models on multimodal data, it is still an open question whether such methods will yield meaningful activity context variables.

For example, in sequence modeling using HMMs, one often does not care specifically about the hidden variables so long as the overall HMM effectively models the observable data. In this case, however, the goal is to find meaningful and interpretable activity contexts since we wish to use them as an organizing principle for investigating language.

5.2 Other methods

Given the possible pitfalls with pure, unsupervised clustering and segmentation methods, it is worth considering other ways to study language in context. One simple idea is to use individual words as the anchor point for activity distributions. For every occurrence of a word, the spatial, temporal and speaker activity measurements can be collected into a distribution which is effectively non-linguistic. That is, a given word induces a distribution over the other modalities. Words can then be clustered by identifying other words with similar distributions by an agglomerative clustering scheme such as distributional clustering (Baker & McCallum, 1998). This method may be useful since the distributions can be manipulated to instead provide distributions over words conditioned on space, time or speaker.

The most direct way to investigate language in the context of activity is to manually label activities. Unfortunately, this is a labor intensive process that may not scale well. Nevertheless, our active team of 10 annotators are still transcribing speech and have direct experience with the data. They may be able to provide good examples of activities in the household as well as tag the activity contexts that occur in their transcription assignments. This incurs a relatively low additional cost to transcription and they have already been "tagging" activities they encounter. Organizing and consolidating these tags into activity contexts is a logical next step for an exploratory analysis. Such human generated annotations are likely to be an important resource.

6 Evaluation and analysis

In its most general form, this thesis presents a study on the relationship between a child's natural home environment and linguistic development. One of the primary ways in which this relationship can be analyzed is by investigating the contributions of environmental variables to early word learning. By correlating predictor variables such as the frequency and recurrence of words in caregiver speech with the child's vocabulary growth, we can quantify the contributions of these variables. Activity context variables, if they are to be of value, should help in modeling and understanding the child's developmental trajectory.

Previous work on the Speechome Corpus linked frequency to word learning (Roy et al., 2009), in agreement with other studies Goodman et al. (2008). Vosoughi et al. (2010) later found recurrence to be more predictive than frequency, and work by Miller (2011) and Shaw (2011) found that spatial context was better correlated with AoA than either of these measures. In this proposal it has been suggested that variables such as recurrence and spatial context may instead be proxy variables for an underlying activity context. Therefore, it is natural to analyze activity contexts as they relate to vocabulary growth. One way to do this is to look at a word's distribution across activity contexts, and see whether contextually "focused" words are learned

earlier, or whether "cross-contextual" words are preferred.

Variables such as frequency are effectively aggregate measures of word's use across the entire 9-24 month period. With activity contexts, word frequency could also be calculated, but restricted to particular activities in which the word's usage is likely to be constrained. We could then study whether particular activities are more conducive to learning. It is tantalizing to wonder whether we could trace back the origin of a word to book reading or playing with a particular toy. This may not be possible, but we would like to know whether language use in one context correlates better to learning than language in another context. The picture that emerged of the child's vocabulary growth in (Roy et al., 2009) showed the child's rate of word learning first increasing at an accelerating rate until about 20 months of age, and then decreasing, leading us to wonder more about the structure and dynamics of vocabulary growth. In the light of activity contexts, could it be that words are strongly contextually tied, and words entering the overall vocabulary are being driven more by some contexts than by others?

Generally, when two words have similar measures for a particular predictor variable but are learned at very different times, we can use these words to probe another predictor variable. For example, if two words have similar frequencies but one is learned much earlier than another, it may be that the difference is captured by another variable. For example, the word "fish" was learned very early despite low overall frequency. From experience, we know that this word held a special meaning for the child since there were colorful magnetic fish on the wall above his crib. The nonlinguistic salience of this word wasn't captured by frequency, but perhaps could be captured by activity context.

Correlating AoA with environmental variables is revealing, but says little about the words the child does not learn. It may be that some activity contexts are primarily oriented around adults, and an interesting question will be whether the language used in these contexts is significantly different from language used in other contexts. For example, it may be that the overlap with the child's vocabulary is generally lower.

There are many quantitative analyses to perform, but there are also important qualitative studies. The general hypothesis is that formats support language acquisition by constraining the topics and interaction dynamics in which language is used. Activity contexts capture the topic of activity, but not the deeper structure, which would require more manual investigation. One way to begin exploring the deeper structure of formats is to simply extract and watch video samples of activity contexts over time. One type of behavior that may be interesting to follow is the development of request. For example, in earlier months caregivers may take most of the initiative during mealtime, but in later months the child may be driving the activity. While we will want to observe examples of activities, analysis of the deep structure of activities is beyond the scope of this thesis. However, this kind of exploration may yield more quantifiable measures that could be operationalized, such as the number of questions the child asks or the number of times the child initiates a conversational exchange.

Before any of the scientific analyses can be performed, the quality of the activity context identification and segmentation must be assessed. That is, do the activity contexts correspond to an interpretable activity? If so, is the data appropriately segmented and labeled? For example, if `diaper-change` really emerges as an activity, what fraction of those labels are correct (precision)

and what fraction of `diaper-change` activities are correctly labeled (recall)? The second question is more difficult, since we do not have an exhaustive labeling of the corpus. Nevertheless, we may try to estimate recall by sampling. This is a technical evaluation, and will be used to characterize which algorithms work the best and how they might be improved. The human generated activity tags may prove useful here. A tool I have written for checking and annotating word births may also be extended to check activity context quality.

7 Conclusion

This thesis investigates early language learning in the dense, naturalistic, longitudinal corpus collected for the Human Speechome Project. The work spans both the technical approaches to working with very large, naturalistic data, and the scientific analysis of one child's linguistic development. The focus of study is early word learning and its relation to environmental "input" factors. With the perspective that language should be treated in context, and motivated by ideas from Jerome Bruner and others, "activity contexts" are proposed. Activity contexts are approximations to *formats*, which are structures that organize the interaction between child and caregiver, support communication and facilitate the child's transition into language. Activity contexts are a critical first step toward investigating formats since they similarly organize the stream of activity into meaningful structures. While they do not necessarily capture the deep structure of formats, they open up the possibility of further analysis since they can be discovered and applied at scale. We suggest that the reference frame provided by activity contexts will be beneficial in investigating word learning. For example, words that are focused on particular contexts may be learned earlier than words spread across contexts. Furthermore, stable activity contexts can be analyzed over time, providing a new picture of language development "in the wild."

The expected contributions of this work fall into several categories. The methodology for large-scale annotation and analysis of naturalistic data may be useful for others working with similar types of data, even if the goal is unrelated to language research. The findings on early word learning contribute to the body of knowledge on child language acquisition. In particular, the language acquisition community may be interested in seeing the patterns of development at both fine granularity and over a long timespan, and the relationship between development and the child's environmental conditions. Finally, by viewing language in context and operationalizing Bruner's notion of formats we hope to better illuminate the link between language, communication and the social world.

8 Research Plan

8.1 Completed work

Although it is an ongoing project, speech transcription is complete for the purposes of the thesis. The vocabulary growth timeline is complete, along with noise robust methods and annotation tools for regenerating and refining the list.

8.2 Timeline

Oct 2011	begin collecting human annotated activity labels
Nov 2011	thesis proposal defense
Nov - Dec 2011	first implementation of activity context discovery and technical evaluation
Jan 2012	clean up and prepare final version of corpus, regenerate frequency, recurrence and other correlations
early Feb 2012	CogSci 2012 deadline, a good target for preliminary findings connecting word learning to activity contexts
Mar 2012	will likely take another pass at activity context discovery methods and technical evaluation
Apr 1, 2012	Interspeech 2012 deadline, a good target for activity context discovery methods
Apr 23, 2012	Media Lab sponsor meeting
May 2012	thesis document outline to Deb
Jun 2012	thesis draft to Deb
Jul 2012	thesis draft to committee
mid Aug 2012	thesis defense
late Aug 2012	submit final thesis document

8.3 Required resources

The critical resources required to complete this thesis are access to the Speechome data servers and our team of data annotators. I have helped to manage these resources for the past four years and do not expect any significant changes or new requirements. In addition, I will use the spatial activity distributions generated by Matt Miller and the person tracker system developed by George Shaw, both of which are available to me.

9 Author Biography

Brandon Roy is a Ph.D. candidate at the MIT Media Lab, studying how children learn language by analyzing behavior in massive audio-video corpora. As a member of the Cognitive Machines group, his work combines cognitive science with machine learning and data mining research. A key element to this research are human-machine collaborative systems that enable people to work with large amounts of data. He is generally interested in the point where perception, action, language and meaning come together for real systems. Prior to MIT, Brandon was part of the research and development team at a silicon valley startup company, working with

Dr. Joy Thomas on machine learning and data mining software for unstructured text corpora. He received his Sc.B. in computer science from Brown University.

References

Baker, L., & McCallum, A. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 96–103).

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*(5), 875–890.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. de, Blei, D. M., & Jordan, M. I. (2003). Matching Words and Pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 127). New York, New York, USA: ACM Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I.(2003, January). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bruner, J.(1983). *Child's talk: Learning to use language*. WW Norton.

Bruner, J.(1985). The role of interaction formats in language acquisition. In J. P. Forgas (Ed.), *Language and social situations* (pp. 31–46). Springer-Verlag.

Dempster, A., Laird, N., & Rubin, D.(1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dromi, E.(1987). *Early lexical development*. Cambridge University Press.

Fleischman, M., & Roy, D. (2007). Unsupervised content-based indexing of sports video. In *Proceedings of the International Workshop on Multimedia Information Retrieval* (pp. 87–94).

Goodman, J., Dale, P., & Li, P.(2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531.

Hofmann, T.(1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50–57).

McCall, R. B.(1977). Challenges to a science of developmental psychology. *Child Development*, *48*(2), 333–344.

Miller, M. (2011). *Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus*. Unpublished master's thesis, Massachusetts Institute of Technology.

Purver, M., Griffiths, T. L., Körding, K. P., & Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* (pp. 17–24). Morristown, NJ, USA: Association for Computational Linguistics.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Cognitive Science Conference.*

Roy, B. C., & Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech.* Brighton, England.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference* (p. 2059-2064). Mahwah, NJ: Lawrence Earlbaum.

Roy, D. K., & Pentland, A. P. (2002, January). Learning words from sights and sounds: a computational model. *Cognitive Science*, *26*(1), 113–146.

Shaw, G. (2011). *A taxonomy of situated language in natural contexts*. Unpublished master's thesis, Massachusetts Institute of Technology.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*(2), pp. 549-565.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition*, *50*(1-3), 431–445.

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

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 306). New York, New York, USA: ACM Press.

Vosoughi, S., Roy, B., Frank, M., & Roy, D. (2010). Contributions of prosodic and distributional features of caregivers' speech in early word learning. In *Proceedings of the 32nd Annual Cognitive Science Conference.*