New Horizons in the Study of Child Language Acquisition[1]

Deb Roy

The Media Laboratory
Massachusetts Institute of Technology

dkroy@media.mit.edu

Abstract

Naturalistic longitudinal recordings of child development promise to reveal fresh perspectives on fundamental questions of language acquisition. In a pilot effort, we have recorded 230,000 hours of audio-video recordings spanning the first three years of one child's life at home. To study a corpus of this scale and richness, current methods of developmental cognitive science are inadequate. We are developing new methods for data analysis and interpretation that combine pattern recognition algorithms with interactive user interfaces and data visualization. Preliminary speech analysis reveals surprising levels of linguistic fine-tuning by caregivers that may provide crucial support for word learning. Ongoing analyses of the corpus aim to model detailed aspects of the child's language development as a function of learning mechanisms combined with lifetime experience. Plans to collect similar corpora from more children based on a transportable recording system are underway.

Index Terms: language acquisition, rich longitudinal data, human-machine collaborative analysis, computational models

1. A New Kind of Data

Language is one of the defining features of the human species, unique in its compositional structure and referential capacity, critical for creation and transmission of cultural knowledge, devastating to an individual when impaired or lost. For all that is at stake, our current understanding of how children learn language is grounded in surprisingly incomplete and biased observational data. As a consequence, many promising theories of language acquisition remain vaguely articulated, contradictory, and untested. More precise and empirically validated theories would shed light on central aspects of human cognition, guide new ways for children to learn, and lead to effective treatment of language disorders.

A critical bottleneck in the study of language acquisition is the quality of naturalistic observational recordings of child development available to researchers. Although young children's language skills change rapidly from day to day, typical naturalistic studies of child development are based on observations spaced weeks or months apart. Sparse sampling leads to a "gallery of before and after snapshots, studio portraits of newborns, and fossilized milestones but little understanding of the process of development itself" [1].

Furthermore, most home recordings of child development consist of speech recordings and/or speech transcriptions but lack any record of non-linguistic situational context. Children of course learn language by connecting words to the people, things, and activities around them. Thus, recording only speech produces an incomplete picture. Although researchers are increasingly likely to complement audio with video recordings, the amount of video recorded tends to be exceedingly sparse due to the cost of analyzing video, and due to the disruptive observer effects of introducing video recording into home environments.

Four years ago my colleagues and I launched the Human Speechome Project with the goal of making a comprehensive and unbiased record of one child's (my son's) development at home [2]. The name of the project has two interpretations. First, our aim is to study speech in the context of the home, hence the combination of "speech + home" to yield the invented term "speechome". Second, this kind of data provides a basis for studying the environmental complement of genetic influences on language development, hence the naming parallel to the Human Genome Project.

We have completed the recording phase of the project yielding the Speechome corpus of approximately 90,000 hours of video and 140,000 hours of audio recordings spanning my son's life from birth to age three. Observational records of this magnitude are now possible due to the ease and affordability of technologies for digital data capture and storage.

The nature of recordings in our study has raised a variety of engineering, design, and privacy challenges. These have been addressed to a sufficient enough degree that I now believe ecologically-valid densely sampled observations of this kind will become pervasive in the study of child development and other areas of human science and design.

The successful completion of the recording phase of the Speechome project has motivated the development of new ways to analyze and interpret large audio-visual corpora. We are developing a human-machine collaborative methodology that enables fast yet accurate speech transcription and video annotation. Building on this method, we aim ultimately to uncover principles of language acquisition through computational models that are grounded in human data.

In this paper I bring together elements from a number of our previous publications to provide a coherent synthesis (hence the high self-citation count that I hope the reader will pardon). I will provide some historical perspective on the origins of the project, report progress on development of analysis tools, preliminary analysis results, and sketch plans for work ahead.

2. Stepping into the Shoes of a Child

A fruitful way to study human cognition is to build machines that "step into the shoes" of humans and perform selected human functions in human-like contexts. This approach forces us to take the machine's point of view and build up whatever mechanisms are needed to perform the target function. The implementation of mechanisms in humans and machines will

1 Invited keynote paper, Proceedings of Interspeech 2009.

of course differ, yet the mechanisms may share functional principles that are easier to discover through the design of working models rather than analysis of more complex natural systems. By analogy, the principles of aerodynamics such as lift and drag that underpin current explanations of bird flight were discovered – as I understand it – through the trial-and-error iterative process of aircraft wing design. Similarly, principles of cognitive processes in humans may emerge by designing cognitive systems and testing them in naturalistic environments.

I originally stumbled into this method about 15 years ago in what began as an engineering effort to build a robot that could learn language in human-like ways. Eventually the tables were turned and I used the robot to study child language learning [3,4]. My original motivation was dissatisfaction with contemporary AI methods of semantic representation. All of the established approaches for representing meaning by machines used networks of word-like symbols, leading to systems hopelessly trapped in circular definitions. Inspired by what was known of early child language acquisition, I developed a robot that could learn from "show-and-tell" interactions with a human teacher. Given a number of visual presentations of objects paired with spoken descriptions of the objects, the robot learned to (1) segment continuous speech in order to discover spoken words units, (2) form visual categories of object shapes and colors, and (3) learn semantically appropriate associations between speech labels and visual categories.

Two key principles governed the robot's learning algorithm. The first was sensitivity to temporally local recurrence structure of both the visual and speech input streams. The second was sensitivity to cross-modal mutual information. These two learning biases were coupled with short and long term memory systems in the robot. Recurrence analysis operated on the contents of a short-term memory system that buffered the most recent few spoken utterances and visual scenes of the robot's input stream. Recurrent speech-visual tokens were deposited in a long-term memory. At longer time scales, mutual information analysis selected semantically relevant speech-to-visual-category associations for placement into the robot's acquired lexicon and also drove a "garbage collection" process to purge semantically inappropriate contents of the long-term memory.

I was able to teach the robot a small vocabulary of shape and color words by show-and-tell. This result, however, did not seem significant since as the designer of the robot I of course knew how to interact with it in order to get it to learn. I could ask others to interact with the robot as a form of evaluation but essentially the same problem would arise – the robot's performance would depend to a great degree on how well I coached others on how to interact with it (this problem seems to plague most work in human-robot interaction). And so the idea was born to use child data to evaluate the robot. If the robot could learn from the same audio-visual input as a child, then the robot's perceptual processing and learning algorithms would demonstrate an interesting level of capability.

To test the robot, my colleagues and I made video recordings of six mothers and their pre-verbal infants as they played with a variety of toys and everyday objects. Audio recordings of the mothers' speech were fed into the robot, aligned with images of the objects that the children were playing with as they heard the speech. With this simplified form of visual input (the robot only saw one object at a time) paired with the child directed speech, the robot was able to learn a small vocabulary of words such as "ball" and "doggy" grounded in visual categories.

Although the original intent of this experiment was to evaluate the robot, the more interesting implications of the results turned out to be their bearing on the nature of child language acquisition. The fact that the robot learned from naturalistic child data provided a proof point that the principles of learning embodied by the robot – a sensitivity to temporally local recurrence and global cross-modal mutual information – were a viable strategy that a child may also use to learn words. The robot served as a new kind of instrument for studying the word learning environment of children. To a limited degree the robot stepped into the shoes of six children and let us evaluate a computationally precise theory of word learning. Whether or not children actually learn according to these principles remains an open question.

Three limitations of this early experiment have motivated the Speechome project. First, lexical semantics in the robot was grounded in perceptual categories yet many basic conceptual distinctions such as an object versus its properties cannot be represented in strictly perceptual terms. For instance, the robot could never in principle learn the difference in meaning of "ball" versus "round" since both terms would be grounded in terms of the same perceptual category. Clearly even a young child has a far richer grasp of word meanings that encompasses not only perceptual categories but also conceptual knowledge of actions and expected outcomes, pragmatic conventions governing word use, syntactic roles, and so forth. To complicate matters, some of the most frequent words in a young child's lexicon include indexicals ("that"), self-reports ("uh-oh"), "good", "yeah", "no", and other words that beg for a richer semantic/pragmatic framework than mere perceptual categories. This realization spawned a line of research on richer models of embodied/situated meaning that focus on action and affordances [5-8]. Although I will not delve further into those models here, I expect they will eventually shape our approach to modeling language acquisition based on the Speechome corpus.

Two further limitations of the early work regard the quality of the observational data. The recordings were made in a child observation lab, not the natural context of the home. Both mothers and children are known to act dramatically differently in novel contexts especially where observers are so pointedly present. To make the recordings, mothers were provided with toys and asked to play naturally. But play makes up only a small fraction of everyday life at home. Moreover, each mother-child pair was recorded for two one-hour sessions a day or two apart. Thus, we only had a snapshot of each child as opposed to longitudinal data that could support the study of language development.

3. The Speechome Corpus

Motivated by the limitations of the earlier experiment, we launched the Human Speechome Project. Before the birth of my first child, my home was outfitted with fourteen microphones and eleven overhead omni-directional cameras. Audio was recorded from ceiling mounted boundary layer microphones at 16 bit resolution with a sampling rate of 48 KHz. Due to the unique acoustic properties of boundary layer microphones, most speech throughout the house including very quiet speech was captured with sufficient clarity to enable reliable transcription. Video was also recorded throughout the

home to capture non-linguistic context using high resolution fisheye lens video cameras that provide a bird's-eye view of people, objects, and activity throughout the home (Figure 1). Recordings were made from birth to the child's third birthday with the highest density of recordings focused on the first two years (I refer to my son in this context as "the child" reflecting the objective stance towards his development that the Speechome corpus enables me to take).

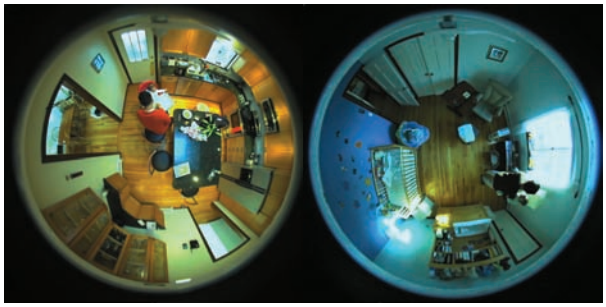

Figure 1: *Sample video frames from the kitchen and baby bedroom.*

With an initial focus on early language acquisition, our current analysis efforts are on the child from 9 to 24 months of age. For this age range, the corpus contains 4,260 hours of recording time spread over 444 of the 488 days (i.e., no recordings were made on 44 of the days across the 16 months) for an average 9.6 hours of recordings per day. We believe these recordings capture approximately 70-80% of the child's waking hours in the 9-24 month period of life. Audio was captured as 14 parallel tracks from each microphone placed around the home. The number of video tracks depended on which cameras were turned on. Typically 5-6 cameras (main living spaces including kitchen and child's room) were active at any given time.

Similar to previous longitudinal case studies, conclusions about the general nature of language development that may be drawn from analysis of the Speechome corpus are inherently limited since the data charts only one child's development. However the corpus differs from previous case studies in important respects. In contrast to diary studies, which are necessarily theory-laden (since the diarist cannot record everything, he/she must rely on theoretical biases to decide what is noteworthy at the time of observation), the Speechome corpus may be re-analyzed multiple times guided by different theoretical perspectives. The existence of high-resolution video provides opportunities to study the role of various aspects of non-linguistic context from joint attention to routine activities and beyond.

4. Analysis / Modeling Framework

Our guiding framework for analyzing the Speechome corpus is to develop computational learning models that, in a limited sense, step into the shoes of the child and sequentially "experience" what the child experienced. Processing is divided into two layers, perception and learning (Figure 2).

The role of perceptual processing is to extract streams of meta-data from audio and video recordings that encode various features relevant for situated language analysis. From audio we plan to extract who was saying what and how (e.g., word level speech transcription, speaker identification, prosody features). From video we plan to encode who was where (person tracking and identification), what were they doing and how (activity classification, manner analysis), and with what objects (object tracking and classification).

The output of the perceptual processing layer feeds into a machine learner that embodies a computationally precise hypothesis of child language acquisition. The output of the learning system may be treated as predictions of what a child would learn. These predictions can be compared with what the child actually did to evaluate the learning system, and thus the viability of the underlying principles of the machine learner as being those used by the child.

Perception and learning are not as cleanly separable as the figure suggests since acquired knowledge may have "top-down" influences on perception. We treat this framework as a general guide for stages of analysis but are ready to admit interactive influences between stages in the course of our research. A more complete framework would thus be diagrammed to include an arrow from the learner layer back to the perceptual layer, however Figure 2 more accurately reflects our current plans.

A fundamental limitation of the framework is its implication that language learning wholly consists of passively processing sensory input. Clearly this is not the case. Language acquisition is an interactive process and I believe the most promising way to model / explain meaningful language use is in terms of interactive processes (e.g., see [5,6]). The passive nature of the analysis framework reflects the inherent limitations of working with "dead data" – frozen records of human interactions. The limitations of observational methods can be complemented by experiments that involve interventions in a child's language learning environment. The Speechome methodology may lead to new types of intervention studies that are embedded in ecologically valid contexts.

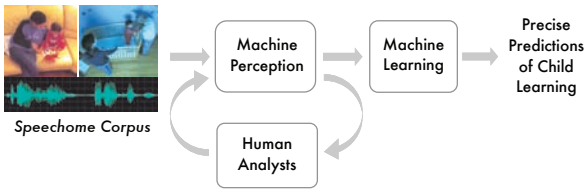

Figure 2: *Analysis framework: Machines that step into the shoes of a child.*

We are working towards instantiations of the complete framework that grounds cross-modal machine learning systems in both audio and visual data streams. However, most of our progress to date has been on the perceptual processing layer of the modeling framework leading recently to preliminary insights into speech patterns.

5. Human-Machine Collaboration

The scale and open-ended nature of the Speechome corpus creates serious challenges for analysis. On one hand, the sheer volume of data renders purely manual methods using currently available tools prohibitively expensive. On the other hand, spontaneous conversational speech recordings and video recordings of cluttered everyday life with uncontrolled lighting result in very high error rates using state-of-the-art technologies for automatic speech recognition, visual object tracking, and so forth.

Our approach to perceptual processing is to combine complementary strengths of human analysis and automated algorithms. We have selected speech transcription and visual person tracking as two key tasks for enabling analysis of the corpus. Speech transcripts will serve as a natural index into the video corpus given the focus on language acquisition. The location and identity of all people provides a basic encoding of both social and activity context as we shall see later. We plan to transcribe all words heard and produced by the child from age 9-24 months – an estimated 10 million words – and to annotate the location and identity of all people in the child's vicinity over the same 16-month period (200 million frames of video). To achieve these goals in a cost effective way, we have developed human-machine collaborative systems for fast speech transcription and video annotation. A third tool called TotalRecall, provides a global view of the corpus contents with support for limited types of speech transcription and video annotation. Chronologically, TotalRecall was developed first so I will describe it first.

5.1. TotalRecall: Audio-Visual Browser

TotalRecall is an audio-video data browser and annotation system [9]. As shown in Figure 3, audio is visualized with spectrograms. Video is visualized using a technique that highlights movement while suppressing static areas of the scene. Users can select which of the 25 "channels" of audio and video to view, and can change the time scale to view periods of data ranging from seconds to years.

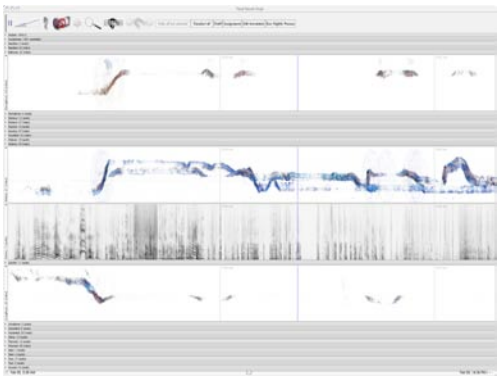

Figure 3: *Screen shot of TotalRecall interface showing approximately two minutes of data across three video and one audio channel. Blue and brown streaks indicate visual movement of people within and between rooms of the house.*

TotalRecall is used in two ways. First, it provides a global view of the corpus indicating when and where recordings were made, and the gist of recording contents as indicated by the presence of speech and movement activity. Second, the system supports limited types of speech transcription and video annotation.

Speech transcription for the project was initially done using TotalRecall. A transcriber who also was a caregiver of the child used TotalRecall to transcribe a substantial portion of the child's early speech as he transitioned from babble to first words – a transcription task that can only be performed by caregivers familiar with the child's idiosyncratic early speech patterns. Speech transcription of more mature speech is now done by a group of transcribers using a faster tool described below. TotalRecall is also used to annotate the room location of the child. This can be done rapidly and precisely with the aid of the video visualizations. Annotations of the child's location are used to select which audio channels are analyzed. Only speech occurring in the room of the child is selected. We call the speech contained in this subset of the data *child-available speech* (CAS). In contrast to child directed speech (CDS), CAS includes not only CDS but also speech between adults that happens to occur in the presence of the child.

5.2. BlitzScribe: Fast Speech Transcription

We performed tests of automatic speech recognition on sample Speechome recordings and found the word error levels to be unacceptably high (well over 75%). Multiple causes led to such poor performance including far-field acoustic recording conditions, spontaneous speaking styles rife with overlapping speakers, and widely variable articulation and prosody ranging from barely articulated coordinating speech among adults preparing a meal together to hyper-articulated parentese. Instead of relying on automatic speech recognition, we have used automatic speech processing to help humans transcribe more efficiently.

BlitzScribe is a tool for accelerating manual speech transcription by a factor of four to six fold compared to other available transcription methods [10]. The transcription process for unstructured free-running audio recordings may be divided into four iterative steps: (1) Find speech (in the Speechome corpus we estimate that 25% of recordings contain speech); (2) Select a segment of the speech that is to be transcribed; (3) Listen to the segment; (4) Type what was heard, and repeat. Using currently available transcription tools a significant portion of time is spent on Steps 1 and 2, and the coordination of these steps with Steps 3-4. BlitzScribe automates Steps 1 (speech detection) and 2 (speech segmentation) and feeds the stream of sound bites into a transcription interface designed for speed. The speech segmenter uses pause structure to find split points within speech that typically occur at word boundaries, and is tuned to produce sound bites that are usually short enough to transcribe after hearing only once because they do not overwhelm the transcriber's working memory (needing to listen to a speech sample multiple times is a major cause of transcription slowdown).

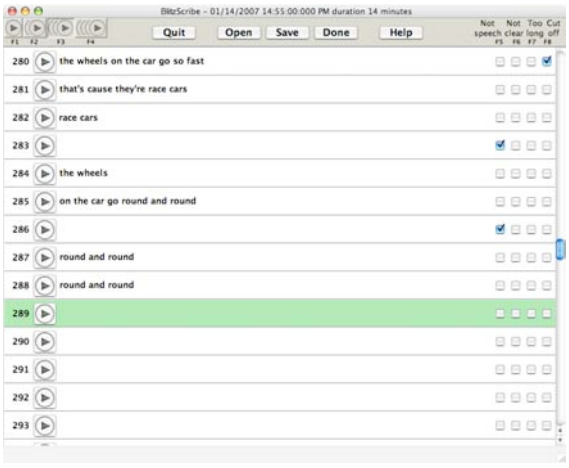

Figure 4: *Screen shot of the BlitzScribe interface with samples of actual speech transcriptions.*

The BlitzScribe user interface (Figure 4) presents each sound bite as a separate row in a scrollable list. Pressing the play button on a row causes that speech segment to be played. At any time the transcriber can begin typing what they hear. When the return key is pressed at the end of transcribing a

sound bite, BlitzScribe plays the next sound bite automatically. The speed of playback effectively tracks the speed of the transcriber. In typical use, the transcriber enters into a "listen-and-type" work flow akin to a stenographer.

In evaluations, BlitzScribe is at least four times faster than the next best speech transcription system without sacrificing accuracy [10]. Using the best competing transcription tool that we could find, one hour of recordings takes about 6 hours of labor. In contrast, the same transcriber takes 1.5 hours using BlitzScribe. Our goal is to transcribe 4,260 hours of audio covering the 9-24 month age period. We project a reduction in transcription time from 25,600 hours to 6,400 hours.

5.3. Speech Processing Pipeline

Transcription using BlitzScribe is embedded in a processing pipeline with the following steps:

1. Channel selection: The audio channel with the highest persistent energy is selected as the source for speech analysis.

2. Speech detection: A boosted decision tree classifies 30ms frames of audio as either speech or not speech.

3. Speech Segmentation: Pause-separated stretches of speech frames are grouped to form speech segments (sound bites) using a four-state Markov model.

4. Child-Availability Filter: The subset of speech segments which occurred in the room with the child are selected for further processing. The selection is made using the video annotations of child position generated by TotalRecall.

5. Speech Transcription: Speech segments are transcribed using BlitzScribe. Transcribers are assigned segments in blocks extracted from 15-minute periods of audio. For privacy, the order of blocks assigned to each transcriber is randomized.

6. Speaker identification: Speech segments are automatically identified as either the child, one of the three primary caregivers (mother, father, nanny), or none of the above using a boosted decision tree.

7. Prosody Analysis: F0 is extracted using Praat , and syllable duration is estimated by forced alignment of speech to transcripts using Hidden Markov Model Toolkit (HTK). We plan in the future to generate normalized energy estimates.

To date, we have used this processing pipeline to transcribe 28% of the audio corpus from the 9-24 month period. Of the 4,260 hours, 1,200 hours have been BlitzScribed yielding 3 million transcribed words with associated speaker identity and prosodic features. In addition to the 3M transcribed words, the speech detector found an additional 300K words that our transcribers marked as unintelligible.

5.4. TrackMarks: Fast Video Annotation

TrackMarks is a tool for tracking the location and identity of multiple people and objects across multiple camera zones robustly in spite of partial or complete occlusions [11]. Once a person's speech has been indentified (Step 6 of the speech pipeline) and the person's position has also been visually identified, speech and video can be linked to each other based on identity, opening up interesting cross-modal analyses (e.g., see Section 6). The system integrates an automatic object tracking algorithm into a human annotator's workstation. The video annotation task is divided into three sub-tasks: (1) manual selection and identification of target objects, (2) automatic target tracking over time, and (3) manual correction of tracking failures. These three sub-tasks are coordinated using the TrackMarks interface (Figure 5). The system provides an efficient means of reviewing automatic tracker output to make corrections, and a "subway map" visualization of track data (lower region of interface) to display movement of people across camera zones.

Preliminary tests show that the positions of three to four people in one hour of video with relatively complex activity can be fully annotated in less than two hours by one person. We expect the efficiency to increase with design optimization of the interface and improvement of the underlying tracking algorithm. To date, TrackMarks has only been used for pilot annotation tests. We plan to soon deploy it at scale to fully annotate the positions and identity of all people in the child's vicinity over the 9-24 month age period.

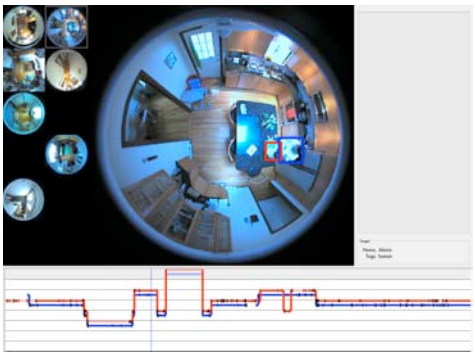

Figure 5: *The TrackMarks user interface.*

There is much more detail available for analysis in the video recordings than mere body position. We have, for instance, developed automated head orientation trackers [12] and are also experimenting with body orientation and gesture characterization algorithms. Over time these additional forms of human activity features along with object tracking and classification will be folded into our analyses.

6. Early Insights into Word Learning

Recently we performed analyses of a 400,000 word subset of speech transcriptions drawn from 72 days that evenly cover the 9-24 month period [13]. Our goal was to gain an initial glimpse into the continuous processes of word learning, and led to some surprising results

6.1. Word Births

We define a *word birth* as the moment of the first reliably transcribed utterance of a new word type by the child. Two caveats need to be made about this definition of a word birth. First, as we continue to fill in transcriptions of the data, the birth of many words will undoubtedly shift forward in time as earlier productions are discovered in the data. Second, it is well known that children comprehend words before they begin using them. Word births as defined here only mark the moment of first transcribed use by the child.

A total of 517 word births were found in the 400K word sample. Figure 6 shows the number of word births binned by month. Although it is widely known that children's vocabularies grow more or less exponentially in this developmental period, we found the rate of worth births abruptly drops at 20 months leading to a "shark's fin" curve. Note that the child's cumulative productive vocabulary size

continues to grow since there are new births each month, but there is a surprising pivot in the curve at 20 months of age.

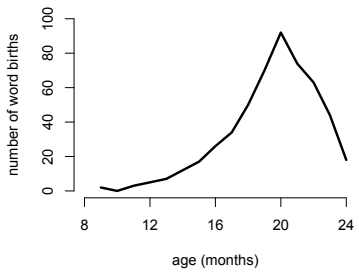

Figure 6: *Number of word births per month over the 9-24 month period.*

We are not certain why this curve has the form that it does. One possibility is that in spite of the rate of externally observed word births, the child's vocabulary is continuing to grow exponentially but due to Zipf's distribution, words learned later are less likely to be observed in productions. The convolution of the exponential vocabulary growth curve and the falling tail of the Zipf distribution lead to the shark's fin form. Another contributing factor may be that as the child discovered the combinatorial power of multiword utterances, he shifted effort from learning and producing new words to putting known words together in new sequences. Further investigations will aim to explain the shape of the curve.

6.2. Caregiver Speech and Word Births

Previous studies have shown that the frequency of a word in child directed speech predicts the age of acquisition of the word by the child [14]. In agreement, we found a significant correlation of -0.29 between the log frequency of words in the child's input and the date of the word birth (Figure 7a), and a stronger and significant correlation of -0.54 for nouns [13].

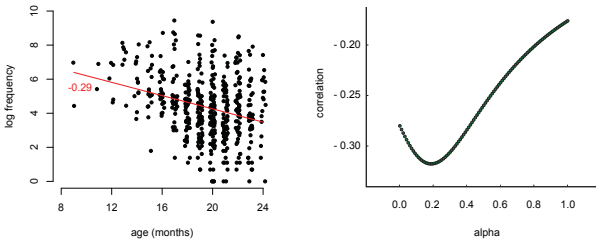

Figure 7: (a) *Words plotted by their date of birth versus the log frequency of the word in caregivers' speech over the 9-24 period. The best linear fit and r-value are shown in red. (b) The combination of how often a word is said and how it is said (based on vowel duration) predicts word births better than either alone.*

The Speechome corpus provides a unique opportunity to study the role of prosody in language acquisition. In addition to how often a word is said, *how* it is said may also affect a child's ability to learn it. In our first look at prosody, we asked whether emphasis placed on words in caregiver speech as marked by syllable duration strengthen our ability to predict word births compared to frequency alone [15]. We assigned each word type a durational emphasis "score" by extracting duration for all vowel tokens, converting these to normalized units for each vowel separately, and then measuring the mean standardized vowel duration for each word type. The log frequency of each word type was combined with the emphasis score using a linear factor, *alpha*. With *alpha*=0 we obtain predictions of word births using frequency alone, and with *alpha*=1 we obtain predictions using durational emphasis alone. As Figure 7b shows, the correlation is strengthened from -0.29 to -0.33 by combining prosodic and frequency measures. This provides modest evidence that the child is leveraging both how often words are said and how they are said in order to learn words.

The method for evaluating the predictive power of two input factors on word births may be extended to include any number of additionally hypothesized factors. In Section 7 I sketch our plans for using this framework for studying the influence of interpersonal distance.

6.3. Caregivers' Fine Lexical Tuning

Vygotsky conceived the zone of proximal development (ZPD) as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers." [16] Vygotsky's view was that the ideal learning environment for a child – the ideal social scaffolding for learning – is to provide experiences within his/her ZPD.

If we view word learning as a kind of problem solving, a period of time leading up to the word's birth might be regarded as the ZPD for that word. A variety of intriguing questions arise with this perspective. Do caregivers adjust the complexity of their utterances that contain a particular word type in a way that is tuned to the word's moment of birth? Might caregivers have a predictive ability in this regard and adjust the complexity of their speech in ways that anticipate word births? The 400K sample of the Speechome corpus has sufficient density to address these questions.

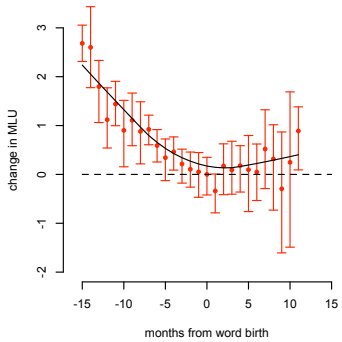

Figure 8: *Change in mean length of one caregiver's utterances in relation to word births. Error bars are 95% confidence intervals.*

We used utterance length, measured in words (not morphemes), as an indicator of caregiver utterance complexity. For each word type that appeared in the child's productive vocabulary by 24 months, we measured the mean length of all utterances each month that contained that word type. The result is a time-varying caregiver utterance complexity curve associated with each word type. The curve for each word type was time shifted so the moment of birth of the word type is aligned across curves. The average change in complexity curves for one of the three primary caregivers is shown in Figure 8. Similar results were obtained for all three caregivers [13].

The result shows that caregivers gradually decrease the length of their utterances containing a particular word type up to the

moment of birth of that word, and then gradually increase complexity. Remarkably, this adjustment of speaking style is tuned to the hundreds of individual word types in the child's productive vocabulary across hundreds of thousands of caregiver utterances. Even more surprising is that the gradual drop in complexity precedes word births by many months suggesting that caregivers have long range predictive abilities.

Evidence of fine lexical tuning of caregiver speech revealed in this analysis raises questions about how and why fine tuning occurs. Perhaps much of the caregiver speech early on is between adults and happens to contain words the child will eventually learn, and thus reflects the complexity of adult-adult speech. As the child enters into the language, caregivers slowly adjust utterances in recognition of the child's growing lexical abilities, bringing utterance complexity to a minimum to meet the child at the moment of birth and gently lifting him into more complex uses of each word type. Or perhaps children learn words from shorter utterances first and are driving the process. Further detailed studies will explore these and alternative explanations.

7. Plans for the Speechome Analysis

In this section I sketch a few planned directions in our ongoing analysis of the Speechome corpus.

We will study the process of grammar development by tracing the formation and transformation of grammatical constructions over time. We are curious to what degree similar semantic and pragmatic contexts predict the use of particular grammatical constructions. The shift in degree of context-boundedness may provide a useful way to study the child's acquisition of abstract grammatical "rules". Of special interest in this regard is the onset of iterative operations including recursion.

The role of joint attention in language acquisition has received significant attention over the past couple of decades (e.g., [17]). The behavioral cues of choice tend to be eye gaze and pointing gestures. Although head orientation (as a surrogate for eye gaze) and gestures may be analyzed in the Speechome corpus, the location of people's bodies provides an equally interesting and important lens into patterns of social interaction and joint activity, and one that has not been studied longitudinally to date.

Figure 9 aggregates the paths of father and child generated using TrackMarks over a 60-minute period in the living room. This visualization reveals two clusters of intense joint activity where the positions of the father and child remain in close mutual proximity for a sustained period of time. These "social hot spots" occur on the couch near the right of the image and near the center of the room on the floor. This structure of interpersonal spatial dynamics leads us to ask: Does the child learn words earlier if they are heard more often in the context of a social hot spot? We plan an extension of the predictive word births analysis (Section 5.2) to answer this question.

Another question guiding our plans is whether the child shows identifiable bodily movements that can be used to detect the *receptive* birth of a word type. For example, it might be the case that once the child learns to understand a new object name (e.g., "ball"), he will often search his environment for a referent soon after hearing instances of the word. Although there are many other factors that will confound any single instance, over thousands of trials (i.e., thousands of times that the child hears a caregiver say "ball") a phase shift in the child's orienting head and body movements might be discernable with the inflection point of the phase shift marking the receptive birth of the word. In cases where the child orients and fixates, we can further analyze the objects in the child's line of sight to verify the presence of semantically appropriate referents. We plan to explore this idea using new computer vision techniques that are tuned to human orienting behavior.

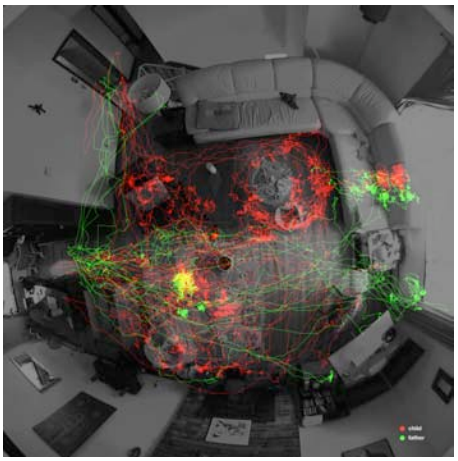

Figure 9: *Sixty minutes of father (green) and child (red) position traces in the living room reveal two social hot spots.*

I believe one of the most promising ways to think about the holistic process of language acquisition is in terms of Wittgensteinian language games as propounded by Jerome Bruner [18]. The basic idea is that a child understands how to participate in games such peek-a-boo without reliance on, and prior to language. Games provides a meaningful context to ground the semantics and pragmatics of words and speech acts. When mother says, "where'd mommy go? Here I am! Yeah!", while engaged in a round of peek-a-boo, the child is able to learn the meaning of these words because of their embedding in meaningful joint activity. We can generalize the notion of a game from literal games such as peek-a-boo to any routine social activity (the game of breakfast, the game of taking a bath, etc.) yielding a framework for analyzing a large variety of naturalistic data. I envision discovering and encoding routine activities in the Speechome corpus using a combination of pattern discovery algorithms and human annotation. Together with the completed transcription of the speech recordings and people's locations, machine learners will process the entire speech stream embedded in corresponding activity contexts from which the machine will learn mappings from words and phrases to semantically and pragmatically appropriate elements of activity structures, including reference to objects and people in the environment. This would become our first complete instantiation of the modeling framework described in Section 3.

8. Beyond N=1

An obvious limitation of the Speechome project is its reliance on one child's data. However conclusively we are able to study phenomena regarding this child, the generalizabilty of results will require data from more children in their natural contexts. With this in mind we have designed a new recording device.

The cost and complexity of the original recording installation was high for two reasons. First, the goal of the design was to conceal all wires and equipment to integrate the system

seamlessly into the home. To do this, over 3,000 feet of concealed wiring was run throughout the house, with holes cut into ceilings to mount microphones and cameras. Second, as a pilot study, we aimed for comprehensive coverage. Thus, every room in the house was instrumented, leading to 11 cameras and 14 microphones installed throughout the home, many of which in retrospect were non-critical.

Figure 10: *The Speechome recorder.*

To reduce cost and complexity of naturalistic longitudinal recordings, our lab has designed a transportable device called the *Speechome recorder* that captures the same quality data as the original corpus from one room, and additionally captures a second video stream from child's eye level to capture details of faces and gestures to augment the birds-eye view. The first prototype of the device, shown in Figure 10, resembles an arching floor lamp. The head of the recorder houses the same model of camera and microphone used in my home. All wiring runs through the mast to the base of the unit, which contains the second camera, disk storage sufficient to hold about 60 days of continuous recordings, computers for data compression, and a touch display controller. The mast can be adjusted to fit into most home settings with a ceiling brace for safety and stability. The Speechome Recorder can be installed, moved, or removed in minutes.

We expect to begin high-density longitudinal recordings in several children's homes in the near future both to increase our diversity of data on typically developing children, and also to begin studies of children with specific developmental disorders that affect communication and social interaction.

9. Conclusions

Preliminary analyses of the Speechome corpus have revealed new insights into the processes of word learning for one child. High density, naturalistic records of child development coupled with appropriate analysis and modeling methods promise to advance our understanding of language acquisition and other aspects of child development in fundamental ways. The field is ripe with opportunities to advance our understanding of language acquisition through cross-disciplinary methods that bring together the human sciences with computational sciences and design.

10. Acknowledgements

I have drawn on ideas from, and joint work with Rupal Patel, Philip DeCamp, Brandon Roy, Rony Kubat, Michael Frank, Soroush Vosoughi, Leo Tsourides, Matthew Goodwin, George Shaw, Stefanie Tellex. The Speechome project is supported by Media Lab industrial consortia, the MIT Center for Future Banking, the NSF, DOD, and ONR.

11. References

[1] Adolph et al. What is the shape of developmental change? Psychological Review (2008) vol. 115 (3) pp. 527-543.

[2] Roy, D. et al. (2006). The Human Speechome Project. Twenty-eighth Annual Meeting of the Cognitive Science Society.

[3] Roy, D. (1999). Learning Words from Sights and Sounds: A Computational Model. Ph.D. in Media Arts and Sciences, MIT.

[4] Roy, D. and A. Pentland. (2002) Learning Words from Sights and Sounds: A Computational Model. Cognitive Science, 26(1), 113-146.

[5] Roy, D. (2005). Semiotic Schemas: A Framework for Grounding Language in Action and Perception. Artificial Intelligence, 167(1-2):170-205.

[6] Roy, D. (2008). A Mechanistic Model of Three Facets of Meaning. In Symbols, Embodiment, and Meaning, de Vega, Glenberg, and Graesser, eds.

[7] Gorniak, P. and D. Roy. (2007). Situated Language Understanding as Filtering Perceived Affordances. Cognitive Science, 31(2), 197-231.

[8] Fleischman, F. and D. Roy. (2005). Intentional Context in Situated Language Learning. Ninth Conference on Computational Natural Language Learning.

[9] Kubat, R., et al. (2007). TotalRecall: Visualization and Semi-Automatic Annotation of Very Large Audio-Visual Corpora. Ninth International Conference on Multimodal Interfaces (ICMI 2007).

[10] Roy, B. (2007) Human-Machine Collaboration for Rapid Speech Transcription. M.Sc. in Media Arts and Sciences Thesis.

[11] DeCamp, P. and D. Roy. (2009). A Human-Machine Collaborative Approach to Tracking Human Movement in Multi-Camera Video. ACM International Conference on Image and Video Retrieval.

[12] Philip DeCamp. (2007) HeadLock: Wide-Range Head Pose Estimation for Low Resolution Video. M.Sc. in Media Arts and Sciences Thesis.

[13] Roy, B., Frank, M., & Roy, D. (2009). Exploring Word Learning in a High-Density Longitudinal Corpus. Thirty-first Annual Meeting of the Cognitive Science Society.

[14] Huttenlocher, J., et al. (1991). Early vocabulary growth: Relation to language input and gender. Developmental Psychology, 27 (1236-248).

[15] Vosoughi, S. et al., (2009, submitted). The Relationship of Input Word Frequency and Prosody to Word Production in Dense Longitudinal Data.

[16] Vygotsky, L. (1978). Mind in Society: Development of Higher Psychological Processes, p. 86.

[17] M. Tomasello and J. Todd. (1983). Joint attention and lexical acquisition style. *First Language*, 43:197–212.

[18] J. Bruner. (1983). Child's Talk: Learning to Use Language. Norton.