The Human Speechome Project

Deb Roy, Rupal Patel*, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak

Cognitive Machines Group, MIT Media Laboratory

*Communication Analysis and Design Laboratory, Northeastern University

Abstract

The Human Speechome Project is an effort to observe and computationally model the longitudinal course of language development for a single child at an unprecedented scale. The idea is this: Instrument a child's home so that nearly everything the child hears and sees from birth to three is recorded. Develop a computational model of language learning that takes the child's audio-visual experiential record as input. Evaluate the model's performance in matching the child's linguistic abilities as a means of assessing possible learning strategies used by children in natural contexts. First steps of a pilot effort along these lines are described including issues of privacy management and methods for overcoming limitations of fully-automated machine perception.

Stepping into the Shoes of Children

To date, the primary means of studying language acquisition has been through observational recordings made in laboratory settings or made at periodic intervals in children's homes. While laboratory studies provide many useful insights, it has often been argued that the ideal way to observe early child development is in the home where the routines and context of everyday life are minimally disturbed. Bruner's comment is representative:

> *I had decided that you could only study language acquisition at home, in vivo, not in the lab, in vitro. 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. We went to the children rather than them come to us.* [Bruner, 1983]

Unfortunately, the quality and quantity of home observation data available is surprisingly poor. Observations made in homes are sparse (typically 1-2 hours per week), and often introduce strong observer effects due to the physical presence of researchers in the home. The fine-grained effects of experience on language acquisition are poorly understood in large part due to this lack of dense longitudinal data [Tomasello and Stahl, 2004].

The Human Speechome Project (HSP) attempts to address these shortcomings by creating the most comprehensive record of a single child's development to date, coupled with novel data mining and modeling tools to make sense of the resulting massive corpus. The recent surge in availability of digital sensing and recording technologies enables ultra-dense observation: the capacity to record virtually *everything* a child sees and hears in his/her home, 24 hours per day for several years of continuous observation. We have designed an ultra-dense observational system based on a digital network of video cameras, microphones, and data capture hardware. The system has been carefully designed to respect infant and caregiver privacy and to avoid participant involvement in the recording process in order to minimize observer effects.

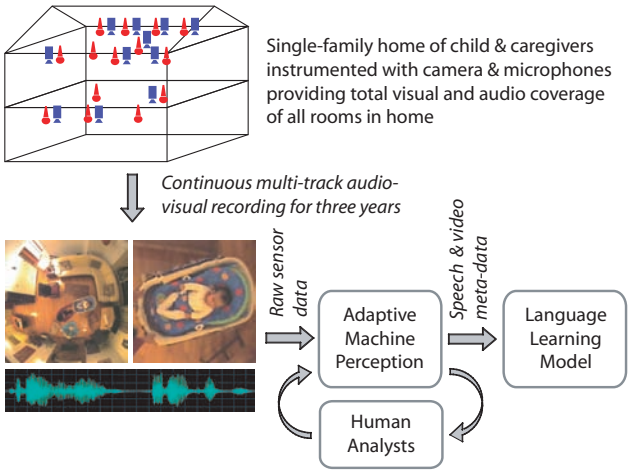

Figure 1: The goal of HSP is to create computational models of word learning evaluated on longitudinal *in vivo* audio-visual recordings.

The recording system has been deployed and at the time of this writing (January 2006), the data capture phase is six months into operation. Two of the authors (DR, RP) and their first-born child (male, now six months of age, raised with English as the primary language) are the participants. Their home has been instrumented with video cameras and microphones. To date, we have collected 24,000 hours of video and 33,000 hours of audio recordings representing approximately 85% of the child's waking experience. Over the course of the three-year study this corpus will grow six-fold.

Our ultimate goal is to build computational models of language acquisition that can "step into the shoes" of a child and learn directly from the child's experience (Figure 1). The design and implementation details of any computational model will of course differ dramatically from the mental architecture and processes of a child. Yet, the success of a model in learning from the same input as a child provides evidence that the child may employ similar learning strategies.

A critical question underlying any model of learning concerns the balance between nature and nurture. HSP brings a new perspective to this age-old debate. Given a near-complete, contextually-rich record of a child's first three years, what are the set of ontological constraints that must be built into a model for it to successfully learn aspects of language? If a machine can be shown to acquire some capability or structure X without corresponding innate preconditions, this provides evidence that the child's environment provides X – and thus need not be innate. While definitive conclusions cannot be drawn from a single subject study, the methodology developed herein enables a new line of investigation that has until now been unexplored.

This paper is structured as follows. Some further background motivation is provided for collecting ultra-dense, *in vivo* observations. The data collection process and a set of initial data visualization and mining tools for semi-automatic annotation are then described. We conclude by sketching preliminary directions of a modeling effort aimed at harnessingthe HSP corpus to examine the role of physical and social interaction in word learning.

Longitudinal *In Vivo* Observation

Bruner's comment is echoed in the work of many researchers who recognize the importance of observing language development as it actually unfolds in the home (e.g., [Bloom, 1973, Nelson, 1974]). The value of longitudinal small-subject studies for examining detailed dynamics of language acquisition is also well established (e.g., [Brown, 1973]). The *density* of data samples is critical. With observations spaced weeks or months apart, apparently sudden changes of linguistic abilities (e.g., new vocabulary, new grammatical constructions) cannot be carefully investigated. Most researchers rely on speech recordings that cover less than 1.5% of a child's complete linguistic experience (e.g., most corpora in the CHILDES archive [MacWhinney, 2000]) – and far less, if any, visual context. As a result, theories of language acquisition hinge on remarkably incomplete observations and are subject to large random effects due to sampling errors. While researchers widely recognize the need for increased observational data to advance empirical and theoretical rigor in the field (e.g., [Tomasello and Stahl, 2004]), few efforts have been undertaken to acquire ultra-dense longitudinal recordings of language acquisition within the home environment.

Dense longitudinal observations, coupled with appropriate data mining and modeling tools, have the potential to make significant theoretical advances. Numerous theories of word learning have been proposed, some positing language-specific learning constraints (e.g., [Clark, 1987, Markman, 1989]), domain-general learning mechanisms (e.g., [Merriman, 1999, Smith, 2000]), or the importance of social inference (e.g., [Baldwin et al., 1997, Bloom, 2000]). There is no obvious way to resolve disputes that arise between these views without a better understanding of the exact nature of children's input as it pertains to each theoretical position. For example, domain-general learning theories may explain results in controlled word learning experiments, but we cannot know which aspects of real-world word learning can be explained by these theories unless they can be tested on more representative data. The same argument holds for the other positions.

Historically, the scope of observational studies has broadened with advances in technology. In the earliest studies, parent-investigators relied on diaries to record observations in the home. Diarists can of course only record a small subset of complete activity – typically first usages and salient errors. Detailed records of input to the child (especially visual and social context) and complete histories of the child's everyday behavior cannot be captured. The introduction of analog and then digital audio recording technology has enabled more detailed and unbiased observation. Video recordings, however, are more difficult to make, are slower to analyze, and thus remain rare in language acquisition studies. Yet, visual context is vital for understanding language development. For example, to study how a child learns the meaning of *thank you*, investigators need to know the non-linguistic contexts in which the phrase was heard and used to understand how the child generalizes from particular instances to new contexts.

In general, many hypotheses regarding the fine-grained interactions between what a child observes and what the child learns to say cannot be investigated due to a lack of data. How are a child's first words related to the order and frequency of words that the child heard? How does the specific context (who was present, where was the language used, what was the child doing at the time, etc.) affect acquisition dynamics? What specific sequence of grammatical constructions did a child hear that led her to revise her internal model of verb inflection? These questions are impossible to answer without far denser data recordings than those currently available.

Ultra-Dense Observation for Three Years

Eleven omni-directional mega-pixel resolution color digital video cameras have been embedded in the ceilings of each room of the participants' house (kitchen, dining room, living room, playroom, entrance, exercise room, three bedrooms, hallway, and bathroom). Video is recorded continuously from all cameras since the child may be in any of the 11 locations at any given time. In post processing, only the relevant video channel will be analyzed for modeling purposes. A sample video frame from the living room camera under evening lighting is shown in Figure 1(left image). The image on the right shows an enlargement of a region of the left image demonstrating the camera's spatial resolution. Video is

captured at 14 images per second whenever motion is detected, and one image per second in the absence of motion. The result is continuous and complete full-motion video coverage of all activity throughout the house.

While omnidirectional cameras provide situational awareness of essentially everything in the house (other than objects occluded due to furniture and people), details such as facial expressions are lost. Although eye gaze and other subtle behaviors are important for language learning, current technology is unable to provide both comprehensive spatial coverage and high resolution. Given our interests in observing and modeling social dynamics expressed in the movements and spatial relations of caregivers and infants throughout the house (see below), we have opted for wide coverage at the expense of resolution.

Boundary layer microphones (BLM) are used to record the home's acoustic environment. These microphones use the extended surface in which they are embedded as sound pickup surfaces. BLMs produce high quality speech recordings in which background noise is greatly attenuated. We have embedded 14 microphones throughout the ceilings of the house placed for optimal coverage of speech in all rooms. Audio is sampled from all 14 channels at greater than CD-quality (16-bit, 48KHz). When there is no competing noise source, even whispered speech is clearly captured.

Concealed wires (above the ceiling) deliver power and control signals to the cameras and microphones, and transmit analog audio and networked digital video data to a cluster of 10 computers and audio samplers located in the basement of the house. The computers perform real-time video compression and generate time-stamped digital audio and video files on a local 5-terabyte disk array. With video compression, approximately 300 gigabytes of raw data are accumulated each day. A petabyte (i.e., 1 million gigabyte) disk array is under construction at MIT to house the complete three-year data set and derivative metadata. Data is transferred periodically from the house to MIT using portable disk drives.

Privacy Management

Audio and video recordings can be controlled by the participants in the house using miniature wall-mounted touch displays (Figure 2, right). Cameras are clustered into eight visual zones (cameras that view overlapping physical spaces are grouped into zones). Eight touch displays are installed next to light switches around the house, each enabling on/off control over video recording in each zone by touching the camera icon. Audio recording can also be turned on and off by touching the microphone icon. To provide physical feedback on the status of video recording, motorized shutters rotate to conceal cameras when they are not recording. The "oops" button at the bottom of the display (marked with an exclamation mark) opens a dialog box that allows the user to specify any number of minutes of audio and/or video to retroactively and permanently delete from the disk array.

Over the first six months of operation, the participants have settled into a stable pattern of use with the privacy controls. On most days, audio recording is turned off throughout the house at night after the child is asleep (typically just after 10pm), and is turned back on in the morning when the child awakes (8am). Audio is paused on average for one hour per day, mainly around adult dinner time, for a total of about 13 hours of audio recordings per day. Video is paused on average 2 hours per day, mainly during nursing, resulting in approximately 12 hours of video per day. The oops button has been used 109 times (63 video deletions, 46 audio deletions) in the first six months of usage. On average, video (audio) segments deleted using this feature have been 7.6 (8.4) minutes long.

Figure 2: Top left: Camera and microphone embedded in ceiling with camera shutter open (the microphone is the small dark disk to the right). Bottom left: shutter closed. Right: Wall mounted touch display for recording control (see section on privacy management for details).

Various other privacy management precautions are being implemented including blackout of video related to bathroom and change table interactions, procedures for retroactive data retrieval from archives upon participants request, and the design of secure data servers to limit access to data.

Handling 338,000 Hours of Data

The network of cameras and microphones are generating an immense flow of data: an average of 300 gigabytes of data per day representing about 132 hours of motion-compressed video per day (12 hours x 11 cameras) and 182 hours of audio (13 hours x 14 microphones). In just the first six months we have collected approximately 24,000 hours of video and 33,000 hours of audio. At this rate, the data set is projected to grow to 142,000 hours of video and 196,000 hours of audio by the end of the three year period. Clearly, new data mining tools must be designed to aid in analysis of such an extensive corpus.

Figure 3 shows a screen shot of a multichannel data browser and annotation system developed for this project. Spectrograms of selected audio channels are displayed in the main window. A fish-eye timeline is used to display large spans of data. The annotator can quickly get the gist of activity throughout the house by viewing all channels simultaneously, and then focus in on

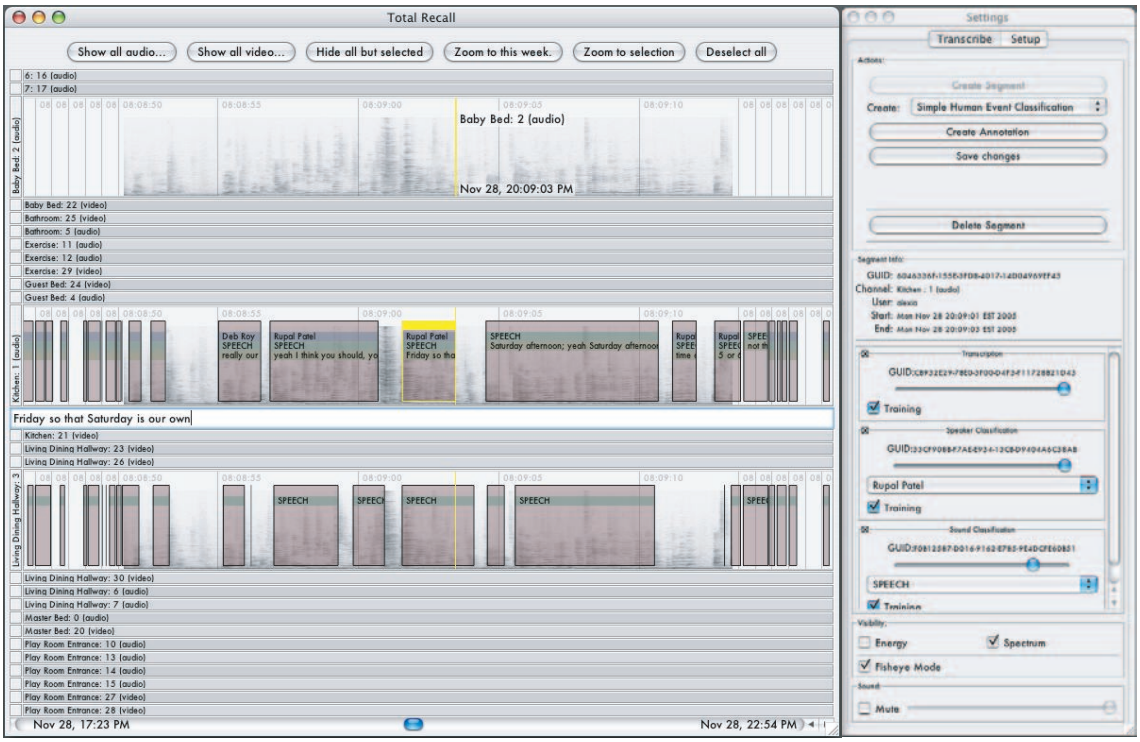

Figure 3: Graphical environment for multi-channel data visualization, playback, and annotation.

regions of interest. To annotate audio, the user selects segments by mouse and uses the right panel to add categorical labels (e.g., speech versus machine noise; identity of speaker, etc.) and text transcriptions. A second window (not shown) displays video from all 11 cameras simultaneously and plays synchronized audio recordings. Future iterations of the browser will include video visualization and annotation tools.

Machine-Assisted Data Analysis

Speech analysis consists of two parallel paths. The first path is to transcribe speech using a state-of-the-art automatic speech recognizer (ASR) designed for speech in noise [Prasad et al., 2002]. Even the best ASR systems will produce high error rates when transcribing unconstrained speech. Thus the second path introduces human annotation. A large corpus of speech which occurs in the vicinity of the child will be manually transcribed in order to obtain a relatively error-free complete transcript of *all* speech heard and produced by the child. The practicality of this latter goal will depend on available human resources for transcription as well as the development of tools to speed transcription. Transcripts will also provide immediate benefits to the development of the ASR, which requires approximately 50 hours of transcribed speech to adapt acoustic and language models for optimal performance in the HSP recording environment.

Current transcription tools are unsatisfactory for working with large, multi-channel recordings. Significant time is spent on finding speech embedded within long stretches of non-speech audio, and in selecting which channel to transcribe given that sound sources usually register on multiple microphones. To address these issues, we have developed a transcription system which automatically finds speech within long recordings using a decision tree algorithm that operates on spectral features extracted from the audio recordings. The speech detection algorithm has been trained using labeled examples created using the annotation system (Figure 3). Regions of speech activity are chunked at pause boundaries from the audio channel with the highest intensity and integrated with a "listen-and-type" interface that automatically paces speech playback to keep up with transcription rate. In initial evaluations, we have found that one minute of conversational speech takes approximately 2.5 minutes to transcribe, signifying a 2- to 3-fold increase in speed over standard transcription tools used in the field.

To automatically generate contextual meta-data for speech transcription, we are experimenting with algorithms for speaker identification, prosodic feature analysis, and audio event classification.

Our long term plan is to adapt and apply computer vision techniques to the video corpus in order to detect, identify, and track people and salient objects. Since the visual environment is cluttered and undergoes constant lighting changes (from direct sunlight to dimmed lamps), automatic methods are inherently unreliable. Thus, similar to our approach with speech transcription, we plan to design semi-automatic tools with which humans can efficiently perform error correction on automatically generated meta-data. The combination of automatic motion

tracking with human-generated identity labels will yield complete spatiotemporal trajectories of each person over the entire three year observation period. The relative locations, orientations, and movements of people provide a basis for analyzing the social dynamics of caregiver-child interactions.

Modeling *In Vivo* Word Learning

As data collection and analysis proceeds, the HSP corpus may be used to study numerous aspects of language including the development of grammatical constructions, prosody, speech acts, and so forth. In this section, we describe first steps of one effort underway to model word learning.

In previous work, we developed a model of word learning called CELL (Cross-Channel Early Lexical Learning) which learned to segment and associate spoken words with acquired visual shape categories based on untranscribed speech and video input. This model demonstrated that a single mechanism could be used to resolve three problems of word learning: spoken unit discovery, visual category formation, and cross-situational mappings from speech units to visual categories. The model operated under cognitively plausible constraints on working memory, and provided a means for analyzing regularities in infant-directed observational recordings.

CELL was evaluated on speech recordings of six mothers as they played with their pre-verbal infants using toys. Recordings were made in an infant lab using a wall-mounted video camera and a wireless microphone worn by the mothers. The speech recordings were paired with video of the same objects recorded by a robot, providing multisensory "first-person" audio-visual input for the model (the video of caregiver-infant interactions was only used to keep track of which toy was in play). CELL successfully acquired a vocabulary of visually-grounded words using a learning strategy that combined within-modality co-occurrence and recurrence analysis with cross-modal mutual information clustering. The model enabled quantitative analysis of the effects of visual context on speech segmentation, and the effects of short term memory size on word learning performance (see [Roy and Pentland, 2002] for details).

Three simplifications made in CELL may be contrasted with our new modeling effort using the HSP corpus. First, CELL was evaluated on a relatively small set of observations. Caregiver-infant pairs were only observed for two one-hour play sessions, held about a week apart. The data was thus a snapshot in time and could not be used to study developmental trajectories. Second, observations were conducted in an infant lab leading to behaviors that may not be representative of natural caregiver-infant interactions in the home. It is unclear whether CELL's learning strategy would work with a more realistic distribution of input. Third, visual input was oversimplified and social context was ignored. The only context available to CELL was video of single objects placed against controlled backdrops. As a consequence, the model of conceptual grounding in CELL was limited to visual categories of shapes and colors underlying words such as *ball* and *red*. It could not learn verbs (since it did not model actions), nor could it learn social terms such as *hi* and *thank you*.

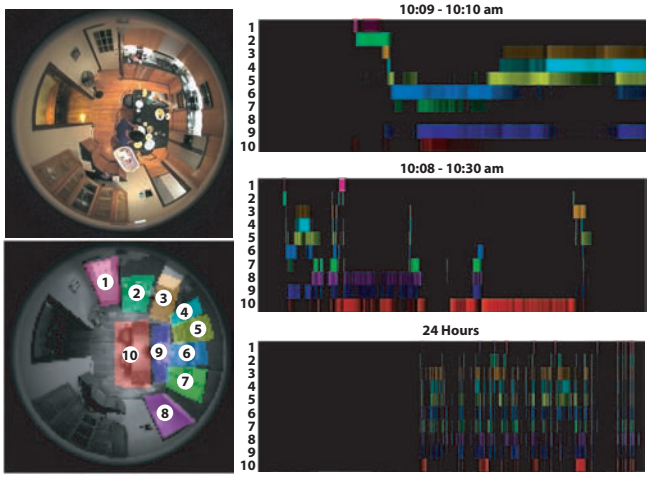

Figure 4: Sample camera image from the kitchen camera (top left), 10 regions of interest (bottom left), and visualization of activity of 1 minute, 22 minutes, and 24 hours (right). For each of the displayed periods, the level of movement in each region is indicated by the brightness of the corresponding horizontal band, with time running from left to right. In the lowest display of a full 24 hour period, three meals are revealed as clusters of activity most clearly indicated in region 10.

The HSP corpus overcomes the limitations inherent in collecting small corpora within laboratory settings. To address the issue of semantic grounding in terms of physical and social action, we have recently developed computational models of perceived affordances for language understanding [Gorniak, 2005] and intention recognition for word learning [Fleischman and Roy, 2005]. In these models, stochastic grammars are used to model the hierarchical and ambiguous nature of intentional actions. In [Fleischman and Roy, 2005], sequences of observed movements are parsed by behavior grammars yielding lattices of inferred higher level intentions. Verb and noun learning is modeled as acquiring cross-situational mappings from constituents of utterances to constituents of intention lattices. We plan to use a similar approach with the HSP data, but with a semi-automatic procedure for learning behavior grammars from video data. Words related to routines (baths, meals, etc.) and names of locations (crib, highchair, etc.) might be modeled on this basis.

The first stage in learning behavior grammars is to identify stable, hierarchically organized patterns of movements that yield a "behavior lexicon". We exploit the fact that we have static cameras in order to divide each room into human assigned regions of interest (indicated as numbered regions on the lower left of Figure 4). These regions correspond to locations or objects in the room that do not move (e.g., the refrigerator) and

have some relevance to various actions (e.g., is used in cooking events). Computing the amount of movement in each region provides a multi-variate representation of movement patterns within a room. This representation is useful for visualizing behavioral patterns at multiple time scales (see right side of Figure 4).

We can recast the problem of learning behavior grammars into one of discovering reliable patterns of movement in these multi-variate data streams. In an initial experiment, we used a set of temporal relations such as *before* and *during* as the basis for learning these patterns. We designed an algorithm that identifies movements in different regions of interest that reliably occur in those temporal relations (e.g. counter-movement before sink-movement). When a relation between such primitive movements becomes significantly reliable, it is treated as a complex movement which itself can participate in temporal relations with other movements. The algorithm proceeds through the data in an online fashion, generating hierarchical patterns of movement until all the data is examined. We have found that these patterns can be used to recognize high level events such as making coffee and putting away the dishes (details forthcoming).

Extensions of this work will focus on developing a video parser that uses grammars constructed from acquired behavior patterns to infer latent structure underlying movement patterns. Cross-situational learning algorithms will be developed to learn mappings from spoken words and phrases to these latent structures.

Conclusions

The Human Speechome Project provides a natural, contextually rich, longitudinal corpus that serves as a basis for understanding language acquisition. An embedded sensor network and data capture system have been designed, implemented, and deployed to gather an ultra-dense corpus of a child's audio-visual experiences from birth to age three. We have described preliminary stages of data mining and modeling tools that have been developed to make sense of over 300,000 hours of observations. These efforts make significant progress towards the ultimate goal of modeling and evaluating computationally precise learning strategies that children may use to acquire language.

Acknowledgments

We thank Walter Bender and the Media Lab industrial consortia for project support, Seagate Corporation and Zetera Corporation for donation of disk storage hardware, Steve Pinker, Brian MacWhinney, and Linda Smith for helpful discussions, and Steve Pinker for naming the project. This paper is based upon work supported under a National Science Foundation Graduate Research Fellowship, a National Defense Science and Engineering Fellowship, and NSF Award SGER-0554772.

References

[Baldwin et al., 1997] Baldwin, D., Markman, E., Bill, B., Desjardins, R., Irwin, J., and Tidball, G. (1997). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, 67(6):3135–53.

[Bloom, 1973] Bloom, L. (1973). *One word at a time*. Mouton, The Hague.

[Bloom, 2000] Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

[Brown, 1973] Brown, R. (1973). *A First Language: The Early Stages*. Harvard University Press.

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

[Clark, 1987] Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In *Mechanisms of language acquisition*. Erlbaum, Hillsdale, NJ.

[Fleischman and Roy, 2005] Fleischman, M. and Roy, D. (2005). Why are verbs harder to learn than nouns? Initial insights from a computational model of situated word learning. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

[Gorniak, 2005] Gorniak, P. (2005). *The Affordance-Based Concept*. PhD thesis, Massachusetts Institute of Technology.

[MacWhinney, 2000] MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

[Markman, 1989] Markman, E. M. (1989). *Categorization and naming in children*. MIT Press, Cambridge, MA.

[Merriman, 1999] Merriman, W. (1999). Competition, attention, and young children's lexical processing. In MacWhinney, B., editor, *The Emergence of Language*, pages 331–358. Lawrence Erlbaum Associates.

[Nelson, 1974] Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review*, 81:267–285.

[Prasad et al., 2002] Prasad, R., Nguyen, L., Schwartz, R., and Makhoul, J. (2002). Automatic transcription of courtroom speech. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 1745–1748, Denver, CO.

[Roy and Pentland, 2002] Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

[Smith, 2000] Smith, L. (2000). Learning how to learn words: An associative crane. In Golinkoff, R. and Hirsch-Pasek, K., editors, *Becoming a word learner*. Oxford Univeristy Press.

[Tomasello and Stahl, 2004] Tomasello, M. and Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31:101–121.