THE BIRTH OF A WORD

by

Brandon Cain Roy

Sc.B., Computer Science, Brown University, 1999
M.Sc., Media Arts and Sciences, Massachusetts Institute of Technology, 2008

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2013

Author ____
Program in Media Arts and Sciences
January 11, 2013

Certified by ____
Deb Roy
Associate Professor
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ____
Patricia Maes
Associate Academic Head
Program in Media arts and Sciences

The Birth of a Word

by

Brandon Cain Roy

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on January 11, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

A hallmark of a child's first two years of life is their entry into language, from first productive word use around 12 months of age to the emergence of combinatorial speech in their second year. What is the nature of early language development and how is it shaped by everyday experience?

This work builds from the ground up to study early word learning, characterizing vocabulary growth and its relation to the child's environment. Our study is guided by the idea that the natural activities and social structures of daily life provide helpful learning constraints. We study this through analysis of the largest-ever corpus of one child's everyday experience at home. Through the Human Speechome Project, the home of a family with a young child was outfitted with a custom audio-video recording system, capturing more than 200,000 hours of audio and video of daily life from birth to age three. The annotated subset of this data spans the child's 9-24 month age range and contains more than 8 million words of transcribed speech, constituting a detailed record of both the child's input and linguistic development.

Such a comprehensive, naturalistic dataset presents new research opportunities but also requires new analysis approaches – questions must be operationalized to leverage the full scale of the data. We begin with the task of speech transcription, then identify "word births" – the child's first use of each word in his vocabulary. Vocabulary growth accelerates and then shows a surprising deceleration that coincides with an increase in combinatorial speech. The vocabulary growth timeline provides a means to assess the environmental contributions to word learning, beginning with aspects of caregiver input speech. But language is tied to everyday activity, and we investigate how spatial and activity contexts relate to word learning. Activity contexts, such as "mealtime", are identified manually and with probabilistic methods that can scale to large datasets. These new nonlinguistic variables are predictive of when words are learned and are complementary to more traditionally studied linguistic measures. Characterizing word learning and assessing natural input variables can lead to new insights on fundamental learning mechanisms.

Thesis Supervisor: Deb Roy
Title: Associate Professor, Program in Media Arts and Sciences

THE BIRTH OF A WORD

by

Brandon Cain Roy

The following people served as readers for this thesis:

Thesis Reader____

Michael C. Frank
Assistant Professor of Psychology
Stanford University

Thesis Reader____

Shimon Ullman
Professor of Computer Science
Weizmann Institute of Science

Acknowledgments

This thesis owes much to the generosity of many colleagues, friends and family. First and foremost, I wish to thank my advisor, Deb Roy[1]. Deb has provided opportunities, support and encouragement throughout. Perhaps most importantly, Deb has shown by example how to combine curiosity, imagination and fearlessness in research. I appreciate the high standard he has held me to, which has always been tempered by his sense of humor. My admiration for Deb extends to his family as well, and I have been fortunate to get to know his wife, Prof. Rupal Patel, and their two children.

I wish to express my gratitude to my committee, Prof. Mike Frank and Prof. Shimon Ullman, for their guidance and feedback, and for bearing with me as this thesis has taken shape. Thanks to Prof. Ullman for his gentle but pointed questions about the deeper implications of this work, which has pushed my thinking in new directions. I particularly want to thank Mike Frank, who has been a collaborator, advisor and friend since 2008. Mike has deeply influenced this work far beyond his role as a thesis committee member, and I am fortunate that our paths crossed when they did.

I came to the Media Lab with the specific idea of stepping back from the disciplines I was familiar with and approaching old questions from a fresh perspective. My hunch was that the Media Lab would be the place to do just that, and indeed, my colleagues in the Cognitive Machines research group have all been exemplars of how to bring creative thinking to hard problems. I would like to thank everyone associated with our group, past and present. In particular, Philip DeCamp and I enjoyed several years as officemates. Philip's a kindred spirit with whom I enjoyed discussing the merits of B-grade sci-fi movies and planning culinary excursions. Philip also built much of the infrastructure that made the Human Speechome Project possible, and combines true artistic integrity with technical genius. Rony Kubat, who for some reason was wearing a red foam clown nose when I first met him, has a sharp wit and a friendly demeanor. Rony was always generous with his time and was a fantastic officemate, collaborator and occasional running partner; you want Rony on your team. Soroush Vosoughi and I still share an office, and somehow whenever we're both there we soon find ourselves in animated conversation and reaching for whiteboard markers. Soroush's energy and good humor are infectious, and I am indebted to him for his feedback,

[1] "No relation!" (our little joke).

critique and encouragement. I would often look to Jeff Orkin to trade notes on "how things work" in the Media Lab. Jeff's good natured take on the complexities of graduate school were always welcome. Michael Fleischman has been a good friend, collaborator, and even employer since I have known him. Stefanie Tellex has been a good friend, collaborator, and even business partner since I have known her. I counted on both of them for their ideas and advice throughout graduate school, and they have always obliged helpfully. I particularly want to acknowledge George Shaw and Matt Miller, both of whom I enjoyed working with and hope we can pick up where we left off again soon. Some of the work in this thesis builds directly on their work. Karina Lundahl deserves a special mention both for keeping our research group functioning smoothly and for partnering with me for the last three years of our speech transcription project. She took care of hiring a great team of transcribers and oversaw many aspects of the project, too many to list here. Karina has been a good friend and a steady presence throughout. To Linda Peterson, thank you for keeping me on track. Thanks to Cynthia Wilkes for her help in the home stretch.

I wish to thank the nearly 70 transcribers that have devoted countless hours of their time and energy to this project. You brought your enthusiasm, interest and ideas to a job that was anything but easy! In particular, I want to acknowledge Halle Ritter who spent a good part of her undergraduate career at Wellesley coming to the Media Lab and working with me. She played a key role in making transcription run smoothly. I also wish to thank Alexia Salata, whose earliest annotation efforts are only a small fraction of her part in this story.

I would like to thank Prof. Brian MacWhinney and Dr. Allen Gorin for serving on my general exam committee. They crafted excellent reading lists and written exam questions. But my oral exam with Deb and the two of them may have been the most intellectually demanding and stimulating 3 hours I have ever had!

I would especially like to acknowledge Dr. Joy Thomas. I worked with Joy right after college up until I came to MIT, and I learned more from him than he may realize. Joy has always been incredibly generous with his time, both when I worked with him at Stratify and since. Joy Thomas, Vineet Gupta, Pandu Nayak and John Lamping were enormously helpful in thinking about and applying to graduate school.

I would be remiss if I did not mention where I did most of my work – the coffeeshops of Cambridge and Boston. In particular, 1369, Voltage, Render, Barrington and Starbucks. There seems to be a coffeeshop renaissance here right now and I have benefitted from it, spending hours at these friendly establishments and consuming countless cups of coffee.

Finally, I wish to thank my family and friends. All of my friends, both old and new, are an amazing bunch and I am lucky to know them. I look forward to many new adventures together. To my parents, Don and Elissa, I only hope that if I have children I can do for them what you've done for Ava and me. You have given us endless love, support and encouragement, leaving us to our own devices but guiding us when we needed it. To my sister Ava, I am continually inspired by you and proud to be your big brother. We have the most fun together! Finally, to Mary – the best thing to come of graduate school is finding you. You are always the brightest part of my day.

Contents

List of Figures

Chapter 1

Introduction

The first two years of a child's life is a period of remarkable growth and development. During this period, children develop motor, social and cognitive skills that will serve them for the rest of their lives. But of all human faculties, the one that is most often noted as being uniquely human is language. How do children acquire their native language? Language acquisition is deeply entangled with perceptual, motor, cognitive and social development. Its roots penetrate deep into biology and culture – it is a product of both nature and nurture. Indeed, the term "language acquisition" is often used in preference to "language learning" since the point where maturational processes end and learning begins is unclear. But few would dispute that language acquisition depends, in part, on a child's environment. What is the nature of this relationship?

In this thesis, we focus on one of the early steps of language acquisition: word learning. Children do not come equipped at birth with the words of their language, but instead they must be learned from the community of language users around them. The words a child hears are part of his linguistic environment, but the linguistic environment is embedded in a larger picture that includes the child's natural, everyday interactions and experiences with objects, activities and caregivers. It is in this rich, multimodal domain that children learn their first words, and so it is in this domain – a child's everyday life at home – that we situate our study of early word learning.

1.1 Thesis overview

The subject of this thesis is early word learning and the environmental contributions to vocabulary growth. Our goal is to investigate this topic through a naturalistic, longitudinal study of one child's productive vocabulary growth, and the ways that his early environment shape this process.

One aspect of this work is descriptive – what does a detailed picture of a child's lexical development look like, what are the characteristics of his early environment, and what is the relationship between the two? But we also approach this subject from a theoretical perspective: that word learning is supported by the rich, naturally occurring interaction structures and activities of everyday life. From this point of view, a child's experience participating in social interaction routines and activities with his caregivers should help to structure his linguistic experience; a particular activity not only constrains what people may *do*, it also helps constrain what they may *say*, simplifying the learning problem.

We have several reasons for suggesting that words more constrained by context may be easier to learn. One is that a "context" is a way of denoting a particular subset of experience, limiting the range of likely actions or referents. A word that is tightly coupled to a small number of contexts may have a smaller scope of possible meanings or uses than a word that is spread across many contexts. Another reason that contextual constraint could facilitate learning is more tied to the learner. Contextual knowledge may prime the learner to expect a certain set of words or referents, possibly focusing their attention. A word tightly coupled to a certain situation or activity could be more reliably incorporated into the child's mental model of that context.

Refining this idea and formulating it in such a way that it can be tested will require further development in the pages that follow. But from this idea, we hypothesize that the degree of contextual boundedness for a word in the child's experience should be predictive of when it is learned, and in particular, that more constrained words in caregiver speech should be learned earlier.

To fulfill both the descriptive goal of characterizing vocabulary growth, the learning environment, and the link between the two, as well as the theoretical goal of studying how contextual constraints might facilitate word learning, we take our study into the place where it all happens: the home of a young child.

1.1.1 Challenges

Studying word learning in the home comes with numerous challenges. How can the natural behavior of child and caregivers be observed without affecting it through the interruptions and intrusions of a researcher's presence, while still respecting the participants' privacy? How can we ensure that the important, often spontaneous moments of a child's early language use are captured? To trace the continuity of development, how can we obtain a longitudinal record that spans a child's first years of life? These questions can be addressed through a novel audio-video recording system embedded in the home that provides near complete coverage and makes dense recording the default – a system that is effectively "always on" while giving privacy controls to the participants. But this solution presents an even greater challenge: how to annotate and analyze the massive dataset that results?

Our study spans the first 3 years of one child's life, recording nearly every day for most of the child's waking hours, resulting in more than 200 terabytes of audio and video. But this much raw data is not particularly useful without appropriate annotations; to study word learning, we need speech transcripts. Transcribing speech is notoriously labor intensive, and without new methods we will not be able to transcribe enough data to fully leverage a comprehensive audio-video record of word learning in the home.

An extensive transcribed corpus of speech in the home can provide the basis for a study of word learning. But speech is not text, and in any annotation task there are bound to be some errors. Assembling a record of the words in a child's vocabulary may be particularly susceptible to such errors. Since much of the work in this thesis depends on an accurate estimate of the child's vocabulary growth timeline, care must be taken that our vocabulary estimate is justified by the data.

The perspective we take in this work is to view language in the rich context of everyday life at home. To approach our descriptive goal of characterizing the child's early environment and its relationship to word learning, how do we characterize the nonlinguistic aspects of the child's experience with language? And more directly related to our notion that contextual constraints support word learning, how do we quantify the way in which the child's exposure to language is structured and constrained by his everyday experiences with people and activities?

1.2 Approach and contributions

The work in this thesis is part of a larger language acquisition research project called the Human Speechome Project (Roy et al., 2006). The Human Speechome Project (HSP) was launched in 2005 by Prof. Deb Roy and his wife, Prof. Rupal Patel, with the goal of studying their newborn son's language acquisition from birth to age three. From the outset, our plan has been to capture as much of the child's early experience as possible by means of daily audio-video recordings. The custom recording system used for the Human Speechome Project was designed to address the challenges mentioned above: respect the participants' privacy wishes, minimize observer effects to capture natural behavior, and provide full coverage of the house and support round-the-clock recording to obtain a complete, longitudinal record of behavior.

As planned, the recording phase of HSP ended in 2008 after the child's third birthday. The resultant collection of audio-video recordings consisted of more than 200,000 hours of audio and video, filled more than 200 terabytes of disk storage, and constituted the world's largest record of one child's early experience.

But a complete audio-video record of everyday life at home is just the beginning. Getting to an analysis of early word learning requires much more than raw data and a number of challenges must be addressed. In this thesis, there are two primary kinds of work: scientific and methodological. Consequently, our contributions also fall into these two categories.

1.2.1 Methodology

Studying language in massive audio-video corpora requires new approaches to data annotation and analysis. A basic requirement for most analyses of spoken language are speech transcripts, but manual speech transcription is notoriously time consuming. With a large-scale audio (and video) record of everyday life at home, manually transcribing speech is a daunting task, while fully automatic methods do not, at present, provide sufficient accuracy for our purposes. Therefore, a substantial amount of the work in this thesis is directed toward a new, semi-automatic approach to transcribing speech and the associated management tools and processes. We refer to our transcription tool, called *BlitzScribe*, as a human-machine collaborative system since the work is divided into complementary tasks performed by human and machine. BlitzScribe, and the ecosystem of tools for monitoring progress, accuracy, and for coordinating the efforts of dozens of human annotators over the course of several years are discussed below. In addition to BlitzScribe, tools for speaker identification, labeling activity contexts, and investigating the child's productive vocabulary are also described.

Large-scale data annotation, in which the final product is the accumulation of work by many people and various fully automatic systems over the course of several years, inevitably leads to annotation errors. It is impractical for multiple people to check over each annotation for accuracy, therefore it is important to account for errors in our analyses. Even in smaller studies in which each piece of data is scrutinized by multiple people, errors can occur and may significantly affect some analyses. In our work, we take great care in obtaining the basic element of our analysis – the timeline of the child's productive vocabulary – since it is particularly sensitive to annotation errors. This challenge is also addressed with a human-machine collaborative approach, by combining automatic statistical methods with focused tools to streamline human review.

Working with a large dataset is qualitatively different than working with a small one. Beyond the inevitable annotation errors that are found in a large-scale dataset, a different approach must be taken when there are thousands, if not millions, of data points. In most cases, the questions we ask must be operationalized so that they can be programmed and performed automatically. For example, many studies of early word learning focus on child-directed speech by adults, but determining whether speech is child-directed often requires human judgment. In our case, we use "child available speech" – speech in the presence of the child – since this can be identified unambiguously. Extracting the contextual aspects of word use is also performed automatically, since it would be overly time consuming and expensive to do so manually. Since our approach is developed in service of our analysis goals, methodological and analysis details are generally interleaved and presented together in the following chapters.

1.2.2 Analysis approach and scientific contributions

The focal point of our analysis is a timeline of the child's productive vocabulary – what words did the child use, and when did he first use them? While it is common to refer to the "age of acquisition" of a word, we coin the term "word birth" to refer to the child's first use of a word in our data. This term fits with the perspective afforded by dense, longitudinal data; vocabulary growth is dynamic and each word has a story, from the gestational period of exposure to the word before the birth to its later uses in the child's spontaneous speech.

Word births begin slowly at first and accelerate dramatically. But what is more surprising than this vocabulary explosion is the subsequent "implosion" – after peaking at 18 months of age, the rate of new words entering the child's vocabulary decelerates faster than it accelerated. However, coinciding with this deceleration is an acceleration of the child's use of new word pairs, suggestive of a smooth transition from lexical learning to grammatical development and overall increase in the child's linguistic expressivity. With densely sampled data, rare words in the child's speech that we might not otherwise observe are detected, revealing a larger vocabulary than expected.

The timeline of word births provides the basis for linking environmental factors to word learning. By considering the exposures to a word prior to when it is learned, the "input" characteristics of a word can be related to its uptake. Measuring the relative contributions of different environmental factors to when words are learned gives one way of assessing their importance, and may help in understanding fundamental learning mechanisms.

Learning, in general, requires both a learning capacity as well as a learnable structure. Our focus is on this structure – the characteristics of a word in the child's input that contribute to its "learnability". To give a preview of our findings, a basic result is that words used more frequently by caregivers are learned earlier. However, a much more predictive factor is a word's *recurrence* – a measure of a word's temporal clustered-ness. Words that are more temporally clustered in caregiver speech tend to be learned earlier. Looking beyond a word's speech characteristics, a word also fits into a nonlinguistic context. We find that words in caregiver speech that are more restricted across contexts are learned earlier. In this case, the contexts we analyze are the activities taking place when a word is used, such as playtime, mealtime, or reading books, and the spatial distribution of where a word is used – is it mostly uttered in a particular chair or used throughout the house? Words used by caregivers across fewer activities or in more spatially localized ways are learned earlier, and these variables are more predictive than both frequency and recurrence. We take this as evidence for our proposal that contextual constraints support learning, and in later chapters speculate on why this could be.

1.3 Summary

This thesis describes a new study of early word learning, built around the dense, longitudinal record of one child's early experience captured in the Human Speechome Project. A naturalistic, longitudinal study such as this presents unique opportunities as well as challenges. Collecting a large-scale, ecologically valid dataset is one challenge; preparing it for analysis is another. We devise systems for data annotation and ways to operationalize our analyses that are appropriate for this kind of "Big Data". Our analyses center on word births, first by looking at the child's vocabulary growth itself and then the environmental factors that are predictive of word births.

This thesis presents new methods and findings in the field of early word learning, but it also draws from a rich history of naturalistic, observational studies of language development.

Prior diary studies have often focused on one or a few children and documented particular aspects of development using pencil and paper supplemented with selected recordings. A common theme in such studies is the descriptive goal of characterizing the phenomenon of interest and identifying patterns in the data. Our study shares these aims, although our data collection strategy is more "theory agnostic" in that we record nearly everything, leaving open the possibility for many kinds of analysis.

Although our data collection approach is theoretically neutral, the questions we ask are not. This work is oriented toward the way that language is grounded in experience and how this relates to word learning. Our underlying hypothesis is that, for a young learner, the rich context of everyday experience provides varying levels of constraint and predictability for word use, and that more constrained words should be easier to learn. Our reasons for proposing this idea are both intrinsic and extrinsic to the learner. Intrinsically, when the learner is in a particular context he may have stronger expectations about the more contextually predictable and constrained words. Such words may also be more reliably incorporated into the learner's mental model of a situation. Extrinsically, words linked to fewer contexts may have a more limited scope and fewer possible associations between word and referent.

Naturalistic, observational data can be complex and difficult to work with, but this complexity is also what makes it valuable. By faithfully capturing the complexity of behavior "in the wild", such data can also capture its underlying logic and structure. The patterns that emerge in naturalistic data can guide researchers toward asking better questions, give new insights into the phenomenon of interest, and pave the way for study in controlled laboratory settings. Such laboratory experiments have made tremendous contributions to our understanding of word learning mechanisms, but to really investigate how these mechanisms operate in their environment we must turn to naturalistic, observational study.

1.3.1 Thesis outline

The remainder of this thesis is organized as follows:

Chapter 2: Continuous Naturalistic Recording of Life at Home – We begin with a brief review of prior naturalistic, observational studies of language acquisition, then introduce the Human Speechome Project and the data collection process.

Chapter 3: Efficient Transcription of the Speechome Corpus – We describe our new method for rapid speech transcription and the process of constructing the annotated Speechome Corpus, the basis for all analyses in this thesis.

Chapter 4: The Child's Productive Vocabulary – We briefly review prior lexical acquisition research, then present our method for identifying word births and the child's vocabulary growth timeline. We consider how statistical models and the increase in combinatorial speech might account for the surprising vocabulary growth curve.

Chapter 5: Environmental Contributions to Word Learning – We consider how environmental factors contribute to word learning. We begin by discussing related work, then examine how three characteristics of words in caregiver speech – word frequency, recurrence, and spatial distribution – relate to when words are learned.

Chapter 6: Language Use in Activity Contexts – This chapter begins by discussing prior work on the role of structured interaction formats in language acquisition. We then introduce activity contexts as an operationalized representation for the daily activities that structure the child's early experience. We describe how they are obtained from data and relate caregiver word use in activity contexts to word births.

Chapter 7: Conclusion – In the final chapter, we summarize our contributions, interpret our findings and suggest future directions for research.

Chapter 2

Continuous Naturalistic Recording of Life at Home

An early account of word learning is given in *The Confessions of St. Augustine* (St. Augustine, 1961):

> "This I remember; and have since observed how I learned to speak.
>
> ...
>
> When they [my elders] named any thing, and as they spoke turned towards it, I saw and remembered that they called what they would point out by the name they uttered. And that they meant this thing and no other was plain from the motion of their body, the natural language, as it were, of all nations, expressed by the countenance, glances of the eye, gestures of the limbs, and tones of the voice, indicating the affections of the mind, as it pursues, possesses, rejects, or shuns. And thus by constantly hearing words, as they occurred in various sentences, I collected gradually for what they stood; and having broken in my mouth to these signs, I thereby gave utterance to my will."

St. Augustine's account is a compelling starting point for our investigation into word learning. He claims that he used social and contextual cues to gradually link words to objects over the course of his early experience, until ultimately he could use words for his own purposes. But this account apparently comes from memory, and as a theory for how children learn words it would require a more careful formulation and tests against real data. Nevertheless, we can take from this passage a perspective that word learning builds on recurring social and linguistic experience.

A version of the above quote also serves to introduce Wittgenstein's *Philosophical Investigations* (Wittgenstein, 2009), beginning with the idea of words as names for things and progressing toward the ways in which words are used. Wittgenstein's concern is largely about the relationship between language and meaning, with the a word's meaning ultimately coming down to its *use*. To begin his argument for this conception of language, he gives the example of a builder and his assistant and a "primitive language" in which uttering an object name (eg. "slab!") invokes the assistant to bring the object to the builder. In such a scenario, "slab" refers to an object, but also its meaning is its use as an imperative. Wittgenstein uses this example to introduce his idea of a "language game" as "consisting of language and the actions into which it is woven" (Wittgenstein, 2009, p. 8e).

That language is interwoven with action and that word learning builds on social exchange and communicative inference are two sides of the same coin. For St. Augustine, learning a word requires inferring the intent of others via nonlinguistic cues; for Wittgenstein, a word's meaning derives from how it is used in an activity or situation. Wittgenstein's work was a landmark in the philosophy of language, but it did not purport to provide a scientific explanation for how children learn language. Nevertheless, the perspective that language must be considered in the rich social and activity contexts in which it is used can inform a scientific inquiry into word learning.

How should this perspective shape a study of language acquisition? One way is by emphasizing not only the outcomes, but also the environmental conditions for language development. This chapter describes the Human Speechome Project, a study of early language development through dense, naturalistic, longitudinal recordings of the first 3 years of one child's life at home. The massive audio-video dataset collected for the Human Speechome Project contains not only a detailed record of linguistic development, but also the linguistic "input" from caregivers and the social, physical and situational context of the child's early experience.

Obtaining a dense, naturalistic, longitudinal record from a child's home for a 3 year timespan presents numerous challenges, and the following pages describe our data collection approach. But while the scale of the Speechome dataset is unprecedented, consisting of more than 200,000 hours of audio and video, naturalistic, longitudinal studies of early language development has a rich history. We begin by describing selected prior work before presenting the details of the Human Speechome Project.

2.1 Naturalistic, observational methods

As with many areas of scientific inquiry, carefully designed and controlled experiments have led to tremendous advances in our understanding of word learning and language acquisition in general. A laboratory setting can support experimentation on young children's word recognition (Jusczyk, 1997) and word segmentation abilities (Saffran et al., 1996), tests of children's "Augustinian" ability to infer a speaker's intended referent (Baldwin, 1991), and studies of the cues children use in generalizing names across objects (Landau et al., 1988; Smith, 2000). There is a vast literature on experimental research in children's early language acquisition, and this research has contributed to much of what we know about the mechanisms involved in development and learning.

But another approach to the subject can also be taken. Rather than conducting experiments in a laboratory, subjects can be observed in their natural environment. These methods attempt to study the phenomenon of interest directly. What is the nature of the child's learning environment and his relationship to it? Naturalistic, observational methods come with many challenges, but they can also complement focused, experimental work.

McCall (1977) recognized both the challenges and the value of naturalistic data analysis to developmental psychology. He claims that "We could learn much from a descriptive survey of the environmental and behavioral landscape before charting our experimental expeditions." Abductive reasoning about observations can lead to new hypotheses that can be tested experimentally. McCall frames issues of naturalistic, observational analysis 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?" Bronfenbrenner (1979, p. 19) is more direct. He states that "...much of developmental psychology, as it now exists, is *the science of the strange behavior of children in strange situations with strange adults for the briefest possible periods of time.*" (italics in the original).

2.1.1 Naturalistic studies of language acquisition

A landmark study in first language acquisition was led by Roger Brown, reported in (Brown, 1973). Brown and his colleagues studied the language development of three children. Researchers visited the children in their homes with portable tape recorders and collected samples of their speech, with regular meetings to discuss and analyze the children's progress. In a collection of essays in honor of Roger Brown, Dan Slobin recalls his own experiences in the early 1960's as a core member of the project, describing the orientation of their research toward questions of transformational grammar (Slobin, 1988).

Brown and colleagues were focused on grammatical development, and his study began when one of his subjects was 18 months of age, the other two at 27 months of age. Later work by Lois Bloom (1973) focused on the earlier developmental period before syntax – the period in which children do not combine words together but instead utter "one word at a time". By means of note taking and selected video recordings of her daughter's 9–22 month age range, as well as samples of speech from three other children, Bloom investigated children's language use prior to the emergence of syntax, but her focus was on the extent of children's knowledge of, and transition into, the use of grammar.

2.1.2 Naturalistic studies of word learning

Before children can combine words together, they must learn some words. One naturalistic study of word learning, reported in Bowerman (1978), focused on the acquisition of word meaning. Bowerman studied how her daughters used words, in particular, how their use of words linked to aspects of a situation, object, or action. Bowerman took notes on the children's word use up to about 24 months of age, supplemented with periodic tape recordings. Like many diary studies, perhaps owing to the practicalities of studying natural, spontaneous speech outside of a laboratory setting, Bowerman's subjects were her children.

Braunwald (1978) studied her daughter Laura's lexical acquisition with an emphasis on acquisition, extension and differentiation of word meaning and the link to situational context. Braunwald kept notes on her daughter's lexical development from 8 months to 2 years, supplemented with audio recordings, and presents data on both the dynamics of Laura's lexical acquisition as well as the context-dependent meaning of words in her vocabulary. Braunwald makes the point that Laura effectively "invented her own system of communication", using a small number of words (52 actively used words by 16 months of age) in concert with situational context to communicate effectively in a wide range of day-to-day situations.

An extensive, longitudinal analysis of 18 children's first words is reported in (Nelson, 1973). Subjects entered the study from 10-15 months of age, and were generally involved for a 1 year period, with the goal of capturing first words to the production of early sentences. As part of Nelson's study, interventions to assess various aspects of children's language were performed. For example, a suitcase filled with toys might be presented to a mother and child for purposes of obtaining an index of children's productive language use – how "chatty" they were. Nelson considers a wide range of issues, from the structure of early vocabulary, to strategies for acquisition, to environmental effects.

Dromi (1987) presents a detailed analysis of her daughter Keren's lexical development in a Hebrew speaking household. Dromi's data collection period spanned from Keren's 10–17 month age range, endpointed by her first word productions to the onset of combinatorial speech. As with many diary studies, the data consist of extensive notes supplemented by audio and video recordings. Of particular interest for our work, Dromi finds that lexical acquisition begins slowly, increases in rate, but then shows a marked decline in word learning rate. Dromi argues that the decline marks the end of a clear one-word stage in child speech, before the emergence of syntax. Dromi also analyzes the extension of word meaning in her daughter's speech – were they used for too broad or too narrow a class of referents, or correctly applied? Also of interest to our work, Dromi notes that many of her daughter's words, for which the referent was unclear, were still situationally relevant. She suggests that contextual knowledge, or a "context-based production strategy" elicited her daughter's use of such words. The importance of context is part of Dromi's argument for the value of case-studies and naturalistic data collection in language acquisition research, although she also emphasizes the challenges in case-studies and the importance of experimental research.

2.1.3 Language acquisition through social interaction

Bruner (1983) makes a strong case for naturalistic, longitudinal methods in studying children's early language. In this work, Bruner recounts his research conducted in England in the late 1970s on two children's linguistic development. Roughly once every two weeks, for more than a year, Bruner and his colleagues visited the children in their homes for an hour, obtaining a half-hour recording of playtime activities. From these recordings, Bruner analyzed the transition from prelinguistic communication into language, arguing that games and routines provided a "scaffold" into language.

Bruner's interest in studying language in the home is motivated by a social interactionist viewpoint, building on ideas such as Vygotsky's zone of proximal development (ZPD) (Vygotsky, 1986) in which caregivers help lead the child from their current level of competence toward their developmental potential. The games and routines of everyday life provided a structure for communication, a stable interaction format in which the prelinguistic child could participate with his caregivers, and the foundation for constructing language. Although Bruner does not emphasize lexical acquisition, he is also not specifically concerned with grammatical development; instead, the focus of this study is on the continuity of communicative development, and what he calls the "language acquisition support system" provided by the rich context of everyday life at home.

2.2 The Human Speechome Project

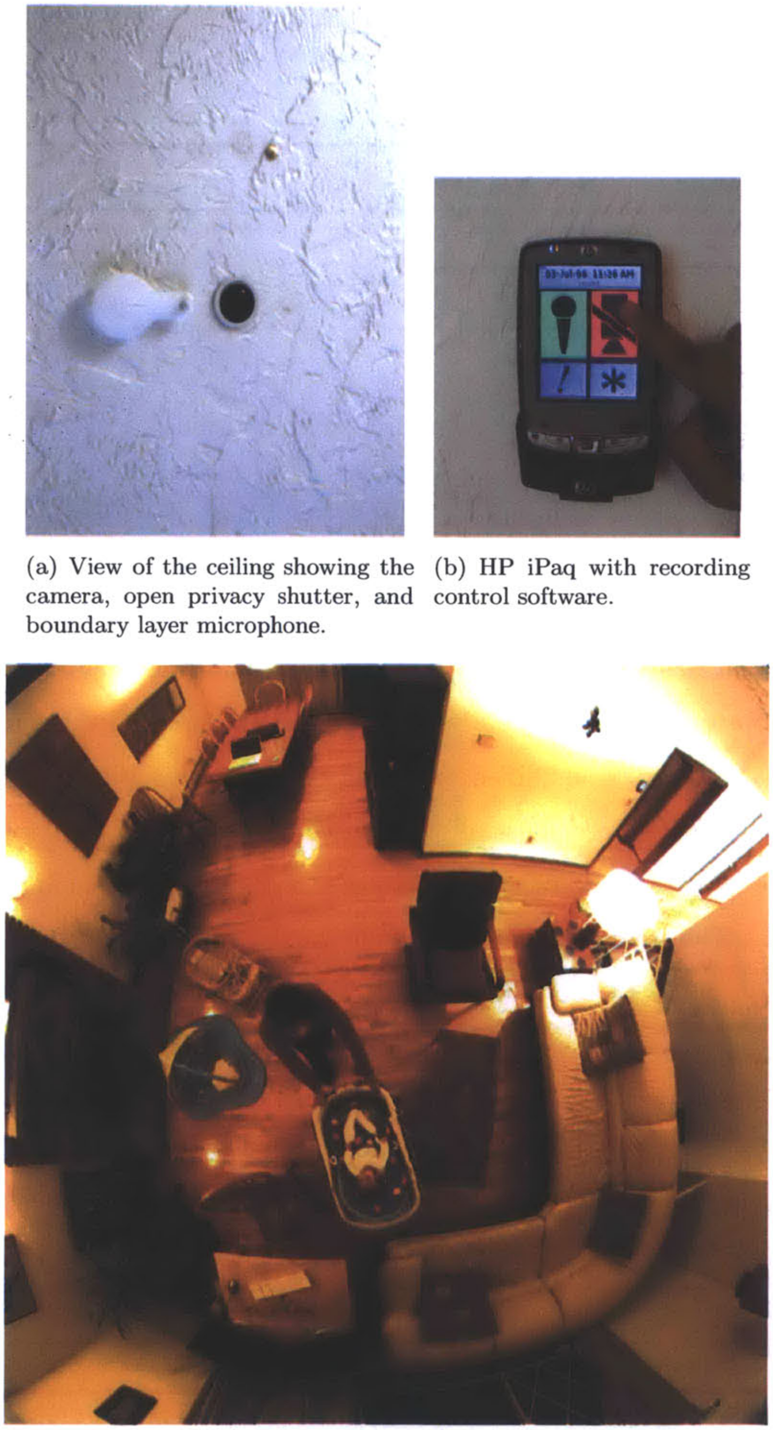
The Human Speechome Project began in 2005 with the goal of studying children's early language development as it naturally occurs in the home. How does a child's everyday experience during his first 3 years of life relate to his communicative and linguistic development? Characterizing daily patterns and the physical and social interactions of early experience, as well as a detailed developmental record, requires dense, longitudinal and minimally obtrusive observational methods. In the summer of 2005, with the advent of inexpensive, high-capacity data storage and high quality audio and video recording technology, capturing several years of multichannel audio and video was within reach. In July 2005, Prof. Deb Roy and wife Prof. Rupal Patel set out to collect the largest-ever corpus of one child's early language development "in the wild", with themselves and their newborn son as the subjects.

2.3 A system for naturalistic audio and video recording

In the summer months preceding the birth of their son, a custom audio/video recording system was designed and installed into the Roy family home. The goal of obtaining full coverage of the child's home and recording from birth to age 3 required a unique and integrated recording system. The system consisted of 11 video cameras and 14 microphones placed throughout the house, a split-level home in the Boston area (see figure 2-1.) The video cameras were embedded above the ceilings, concealing the entire camera body and all attached cables, leaving only the fisheye lens exposed. A movable privacy shutter would cover the lens when video recording was off. Boundary layer microphones, also embedded in the ceilings, were used for audio recording. This type of microphone is affixed to a surface, effectively using the entire surface as a pickup. This yields high audio quality

Figure 2-1: The home of the Roy family and site of the Human Speechome Project.

with a very small footprint – only a small button shaped disc visible on the ceiling. These components can be seen in figure 2-2(a). Within a room, audio volume varies little with position, since distance to the microphone pickup is effectively distance to the ceiling and thus independent of floor position. XLR cables from the microphones and ethernet and power cables for the video cameras and privacy shutters snaked through the ceilings and down to a small server room in the basement. Users controlled the recording system using custom software running on HP iPaqs (a 2005-era touchscreen PDA) that were distributed throughout the house, shown in figure 2-2(b). While all audio channels were controlled as a single unit, video cameras were grouped into zones, and each iPaq controlled video recording within a particular zone. In addition to turning recording on and off, the control software also provided privacy controls for deleting data that was accidentally recorded (the "oops" button) as well as providing a bookmarking function that could be used when something interesting happened (the "ooh" button.) Figure 2-3 shows the layout of the house, with most of the living space and household activity taking place on the second floor. Eight cameras were installed on the second floor (one in each room, and the hallway) and three on the first floor (in the entrance, play room and exercise room.) The bulk of the audio and video recording infrastructure was designed, built and installed by Philip DeCamp, Stefanie Tellex, and Deb Roy in the summer of 2005.

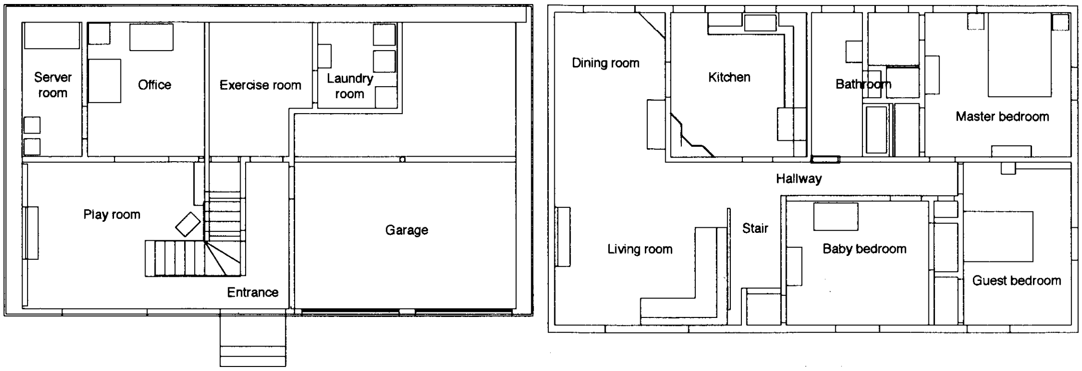


(a) View of the ceiling showing the camera, open privacy shutter, and boundary layer microphone.

(b) HP iPaq with recording control software.

(c) View from the living room camera.

Figure 2-2: Recording and control equipment and a view into the living room.



(a) Lower floor

(b) Upper floor

Figure 2-3: Floorplan of the house. All living space and most activity was on the upper floor.

With this system in place, members of the Roy household began recording their daily lives at an unprecedented scale. Since cameras and microphones were embedded in the ceilings, the system was visually unobtrusive, and recording was easily managed thanks to easy to use software running on readily accessible controllers. Furthermore, the decision to record all day, every day simplified recording management. While the primary reason was to capture as much data as possible, a positive consequence of this approach was that it minimized the effort and attention required of the participants, pushing the system into the background. This reduced the burden the participants and effectively removed the "editorial" component typical of home video recording. In the morning, the participants would turn on the recording system, leave it on all day, and turn it off at night. If the family needed greater privacy at any time they could stop recording or easily delete data that had been recorded. Privacy has been a central concern, and an extensive IRB review process helped in formalizing privacy and data management policies.

2.3.1 Blackeye

Extensive daily recording was ultimately more manageable for the participants and better served the research goals of the project. However, the amount of data collected was significant, with roughly 200 gigabytes of audio and video collected per day. Blackeye, the system responsible for processing the recording commands, capturing and encoding the audio and video inputs, and storing the data to disk was designed and built by Philip DeCamp. This system consisted of a control server that would respond to user commands, video encoding servers for converting the camera outputs to the target video format, a pair of FirePod audio digitizers, and a 5 terabyte RAID array. At 5 TB, the RAID storage would fill up every few weeks, so a tape backup system was used for moving data off the RAID and onto removable 400 GB tapes. All of this was housed in a small server room on the first floor of the house, and approximately once a month the data tapes were transported to the Media Lab to be restored to our local servers.

The underlying data model used by all software layers followed a standard hierarchy, in which recorded data was split into minute long files and stored in a "minute-level" folder, with each audio or video channel producing its own file. The minute-level folders were nested within hour-level folders, then day, month and finally year folders at the top of the hierarchy. Thus, for a given channel and point in time, the data could be retrieved by navigating to the appropriate time and scanning the file for the relevant audio samples or video frames. For example, all audio and video recorded at 5:43 PM on Jan 8, 2007 could be found in the channel-specific files in the `2007/01_jan/08_mon/17/43` folder.

File formats

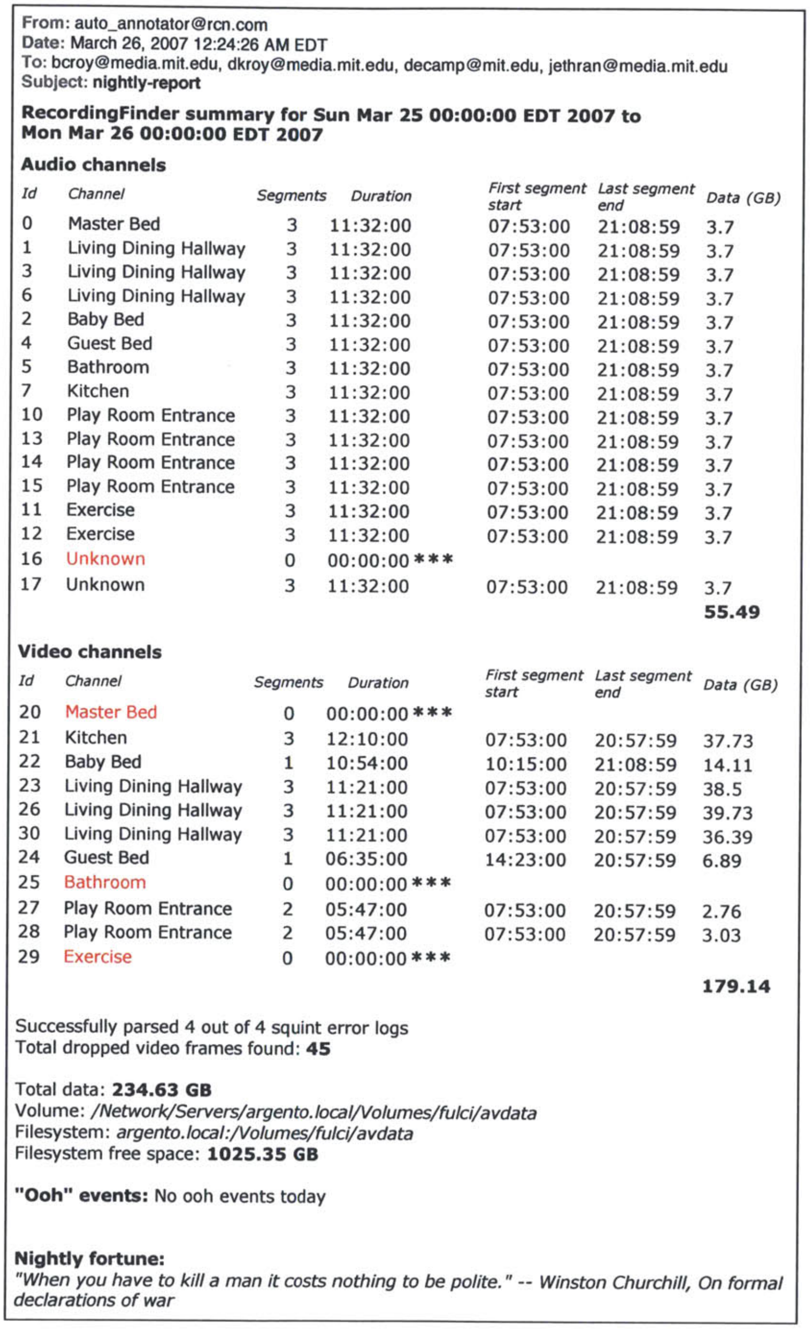
Audio data at the minute level was stored in a custom file format called BEPCM (short for BlackEye PCM.) This is simply a raw, uncompressed PCM audio file with header information containing the audio channel id, start time, end time, number of samples, sampling rate, and so on. All audio was sampled at 48KHz with 16 bit resolution. Video data was stored in a custom file format called SQUINT, with a header describing the video channel id, start time, end time, number of frames, and so on. The video data itself consisted of a sequence of 960x960 pixel JPG images and their associated timestamps. Video was recorded at a variable frame rate, with a maximum of about 15 frames per second when there was motion in the video, dropping to a minimum of 1 frame per second when no motion was detected. This helped to reduce the overall amount of data collected, but SQUINT video files are still significantly larger than than formats such as MPEG that compress across frames. One advantage of this file format is at recording time: processor intensive video encoding is greatly reduced since the cameras supplied a stream of JPEG images. In 2005, realtime MPEG encoding for eleven 1-megapixel video streams was not practical, and even offline encoding would require additional resources. Another advantage of the SQUINT format is at playback time: a single frame of video can be accessed directly without processing adjacent frames. Furthermore, there was some concern about introducing compression artifacts that might interfere with future vision and image processing algorithms. However, a disadvantage is the size of the files generated. After three years of recording, we have roughly 250-300 TB of data, most of which is video. More details on BlackEye and the recording infrastructure can be found in (DeCamp, 2007).

2.4 Data processing and management

The core data collected for the Human Speechome Project is the raw audio and video. Ultimately, all analysis derives from this data, but with roughly 200 GB recorded per day, working with it quickly becomes a significant challenge. To a great extent, raw data is useless without appropriate tools and processes to make sense of it.

One basic challenge is simply keeping track of the recording system. Every evening, when the recorders were typically (but not necessarily) turned off, a nightly recording finder process would scan all data in the filesystem from the previous day and identify the contiguous recording blocks for each channel. These "recording" annotations were stored in a central database and could be queried easily. This also helped in keeping track of data as it was moved from the home to the Media Lab. This information was also emailed nightly to simplify monitoring, as shown in figure 2-4.

To support more efficient browsing of the raw audio and video, a parallel stream of derivative *transform data* was also created. For each SQUINT file, a smaller format, 120x120 pixel "WINK" video was generated online during recording. Spectrogram images for each minute of audio were generated to show the frequency content of the audio signal over time. With



From: auto_annotator@rcn.com
Date: March 26, 2007 12:24:26 AM EDT
To: bcroy@media.mit.edu, dkroy@media.mit.edu, decamp@mit.edu, jethran@media.mit.edu
Subject: **nightly-report**

RecordingFinder summary for Sun Mar 25 00:00:00 EDT 2007 to Mon Mar 26 00:00:00 EDT 2007

Audio channels

Id	*Channel*	*Segments*	*Duration*	*First segment start*	*Last segment end*	*Data (GB)*
0	Master Bed	3	11:32:00	07:53:00	21:08:59	3.7
1	Living Dining Hallway	3	11:32:00	07:53:00	21:08:59	3.7
3	Living Dining Hallway	3	11:32:00	07:53:00	21:08:59	3.7
6	Living Dining Hallway	3	11:32:00	07:53:00	21:08:59	3.7
2	Baby Bed	3	11:32:00	07:53:00	21:08:59	3.7
4	Guest Bed	3	11:32:00	07:53:00	21:08:59	3.7
5	Bathroom	3	11:32:00	07:53:00	21:08:59	3.7
7	Kitchen	3	11:32:00	07:53:00	21:08:59	3.7
10	Play Room Entrance	3	11:32:00	07:53:00	21:08:59	3.7
13	Play Room Entrance	3	11:32:00	07:53:00	21:08:59	3.7
14	Play Room Entrance	3	11:32:00	07:53:00	21:08:59	3.7
15	Play Room Entrance	3	11:32:00	07:53:00	21:08:59	3.7
11	Exercise	3	11:32:00	07:53:00	21:08:59	3.7
12	Exercise	3	11:32:00	07:53:00	21:08:59	3.7
16	Unknown	0	00:00:00 ***			
17	Unknown	3	11:32:00	07:53:00	21:08:59	3.7
						55.49

Video channels

Id	*Channel*	*Segments*	*Duration*	*First segment start*	*Last segment end*	*Data (GB)*
20	Master Bed	0	00:00:00 ***			
21	Kitchen	3	12:10:00	07:53:00	20:57:59	37.73
22	Baby Bed	1	10:54:00	10:15:00	21:08:59	14.11
23	Living Dining Hallway	3	11:21:00	07:53:00	20:57:59	38.5
26	Living Dining Hallway	3	11:21:00	07:53:00	20:57:59	39.73
30	Living Dining Hallway	3	11:21:00	07:53:00	20:57:59	36.39
24	Guest Bed	1	06:35:00	14:23:00	20:57:59	6.89
25	Bathroom	0	00:00:00 ***			
27	Play Room Entrance	2	05:47:00	07:53:00	20:57:59	2.76
28	Play Room Entrance	2	05:47:00	07:53:00	20:57:59	3.03
29	Exercise	0	00:00:00 ***			
						179.14

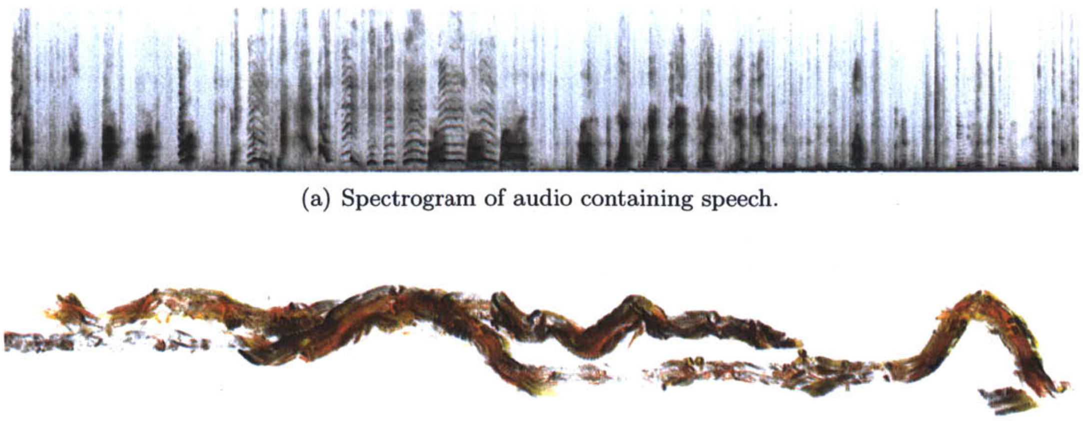
Successfully parsed 4 out of 4 squint error logs
Total dropped video frames found: **45**

Total data: **234.63 GB**
Volume: */Network/Servers/argento.local/Volumes/fulci/avdata*
Filesystem: *argento.local:/Volumes/fulci/avdata*
Filesystem free space: **1025.35 GB**

"Ooh" events: No ooh events today

Nightly fortune:
"When you have to kill a man it costs nothing to be polite." -- Winston Churchill, On formal declarations of war

Figure 2-4: Nightly report email, summarizing the recorded data per channel, total recorded data, errors, and available space on the disk array.



(a) Spectrogram of audio containing speech.

(b) Video volume, showing two people moving.

Figure 2-5: A spectrogram and a video volume, the two basic kinds of transform data for visualizing raw audio and video. A spectrogram shows the frequency content of an audio signal over time. The horizontal axis in the image is time, the vertical axis is frequency, and pixel darkness represents intensity (volume.) A video volume traces out the camera pixels that were changing over time, effectively capturing and visualizing motion in a static image.

a spectrogram, acoustic patterns such as speech can be easily discerned visually. WINK video was processed into "video volumes" (also called "space-time worms") to summarize video content for a minute of video in a single image. A video volume is generated by detecting those pixels that are changing between frames and printing those pixels into an image, but translated horizontally by the temporal position of the video frame. One could imagine stamping a sliding ribbon of paper with the portion of video in motion at every frame, rendering the background transparent while revealing activity patterns. Figure 2-5 shows a spectrogram and a video volume.

As with the nightly recording annotations, video volumes and spectrograms were generated nightly for the previous day's recordings. Each minute of audio or video for a particular channel was processed and saved into a corresponding image file. These transform data images were stored in separate folder hierarchies, specific to the transform data type, but with identical year-month-date-hour-minute folder hierarchies. In addition to the minute-level images, low resolution composite "proxy" images spanning 10 minute, 1 hour and 24 hour durations were also stored for both video volumes and spectrograms. WINKs,

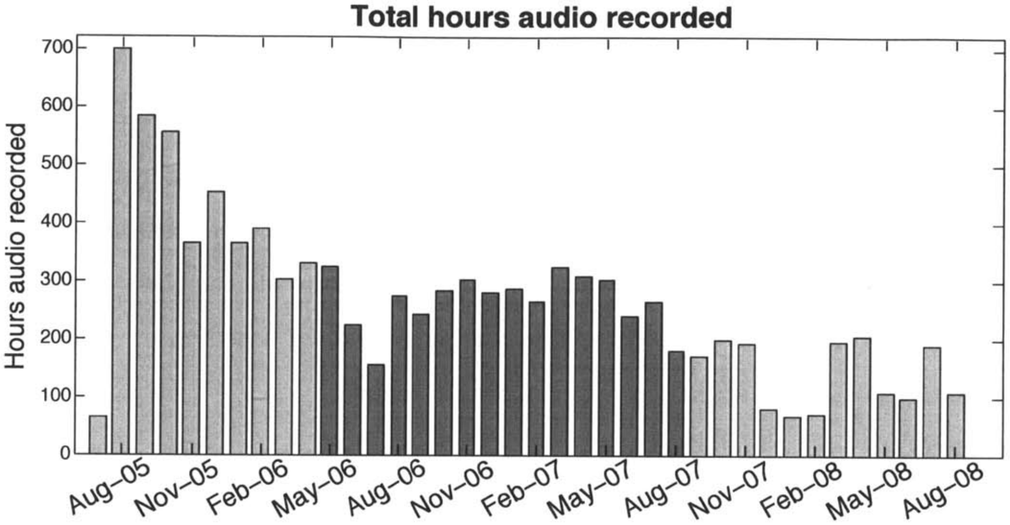


Figure 2-6: Hours of *multitrack* audio recorded by month. Actual amount of audio is roughly 14 times more, since all 14 microphones recorded simultaneously. In the very first months of the project, data was recorded nearly 24 hours a day. The months shown in dark gray are for the child's 9–24 month age range, the focus of the analyses in this thesis.

generated at recording time by BlackEye, were stored in their own parallel folder hierarchy as well.

2.5 The recorded dataset

The full three years of recording for the Human Speechome Project was officially completed at the end of July 2008, although a small amount of data was collected for several months after this period. Figure 2-6 shows the amount of audio recorded by month, with about 64 hours collected in July after the child's birth on July 27, 2005. In the earliest days of the project the participants recorded for nearly 24 hours a day, but as they established a routine they recorded only during waking hours. Altogether, the audio and video recordings consist of roughly 90,000 hours of video and 120,000 hours of audio.

2.6 Conclusion

Language acquisition is an age-old puzzle. Somehow, in the course of their first few years, children begin using words and syntax to communicate. Naturalistic, observational studies have the potential to shed light on this process. Such studies can reveal patterns in the natural dynamics of language development as well as characteristics of the child's early learning environment. Theories can be considered and further developed based on naturalistic data, emergent patterns can better inform experimental research.

Longitudinal, observational studies present many challenges, but from the pioneering work of Roger Brown to the present, they have offered a unique perspective on language acquisition. Our work is based on a new naturalistic, longitudinal study: the Human Speechome Project. An audio-video record of the first 3 years of one child's life was captured by means of a novel recording system in the child's home. The resultant data forms the basis for our work in this thesis. But the raw audio and video collected is not directly usable for our analysis of early word learning. First, appropriate annotations such as speech transcripts must be obtained, as described in the next chapter.

Chapter 3

Efficient Transcription of the Speechome Corpus

Audio and video recordings from the Human Speechome Project are the raw material for our study of early word learning, but they are unsuitable for analysis without appropriate annotations. The basic annotations needed are speech transcripts: what words did the child hear, and what words did he learn? With more than 200,000 hours of audio and video collected over the course of three years of recording, constructing a fully annotated corpus presents a significant challenge. We focus our annotation and analysis efforts on the subset of data spanning the child's 9–24 month age range, since this period typically captures children's first productive word use up to multiword speech (Fenson et al., 1994). Unfortunately, this subset of data is still beyond the capabilities of traditional annotation approaches and new techniques are needed.

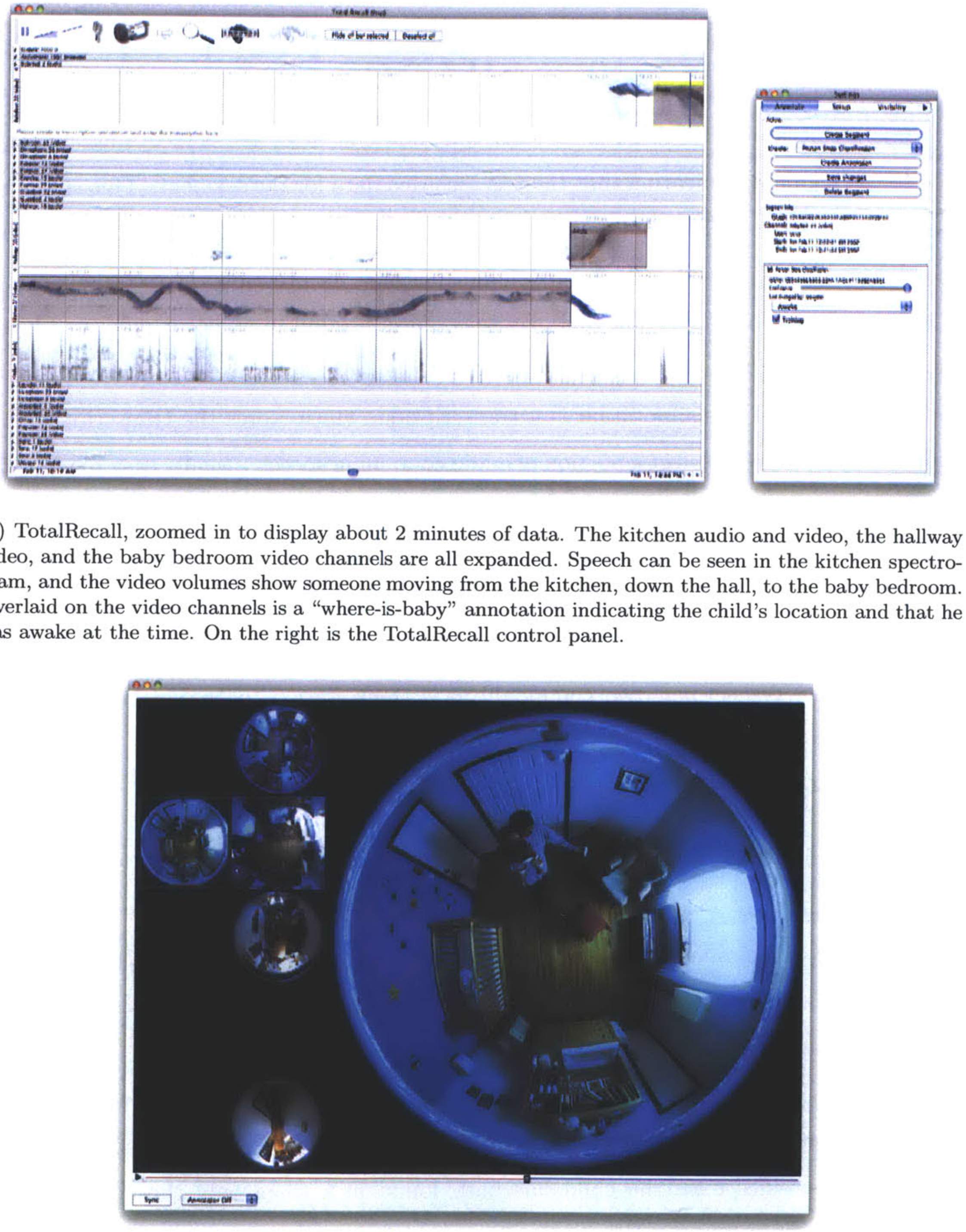
This chapter presents the tools, methods, and processes developed for a large scale annotation enterprise, and concludes by describing the *Speechome Corpus*, the annotated corpus of the child's 9–24 month age range that is the foundation for our study of early word learning.

3.1 Data annotation

At the core of our data annotation process are several tools: *TotalRecall*, *BlitzScribe*, and both automatic and manual methods for speaker identification. BlitzScribe is the pillar that supports our large-scale annotation effort; it is a focused tool designed for rapid speech transcription. TotalRecall is a more general audio and video browsing and annotation tool. To streamline transcription, speaker labeling is performed separately, either with a fully automatic method or with a focused manual annotation tool. This section describes the core elements of our annotation process in greater detail.

3.1.1 TotalRecall

TotalRecall (Kubat et al., 2007) is a data browsing and annotation tool built for the Speechome data. It presents a timeline view of all data across all channels, provides audio and video playback, and supports data annotation. The timeline view for a single audio or video channel displays the spectrograms or video volumes, respectively. As shown in figure 3-1(a), the user interface presents channels as stacked horizontal bands, all time aligned, and the 11 video and 14 audio channels can be minimized or expanded as needed. The user can zoom in and out to inspect portions of the data, with a range from seconds to years. Navigating to a particular point in time is easily accomplished by dragging the timeline to the left or right, effectively scrolling through time. To inspect a particular portion of data, for example, a portion of a spectrogram, the user can highlight the segment and play back the audio. A video volume for a single video channel is useful for determining whether there is activity in the video, and often how many people are present, but it is particularly effective when tracking movement between video channels (ie. when people move from room to room.) The video player, shown in figure 3-1(b), is synchronized with audio playback. For times when a channel was not recording (as determined by the recording finder), no transform data will exist and TotalRecall displays these periods in gray. As an optimization, when there *are* transform data images, the lower resolution 10 minute, hour or 24 hour proxy images are loaded and displayed depending on zoom level.



(a) TotalRecall, zoomed in to display about 2 minutes of data. The kitchen audio and video, the hallway video, and the baby bedroom video channels are all expanded. Speech can be seen in the kitchen spectrogram, and the video volumes show someone moving from the kitchen, down the hall, to the baby bedroom. Overlaid on the video channels is a "where-is-baby" annotation indicating the child's location and that he was awake at the time. On the right is the TotalRecall control panel.

(b) TotalRecall video player, which shows 10 low resolution WINK videos and the selected full resolution SQUINT video. Here we see the child with his father in the baby bedroom, the source of the video volumes displayed in TotalRecall.

Figure 3-1: TotalRecall browser and video player.

When TotalRecall is launched, it requests a username and password for retrieving data from the central HSP database. This database stores metadata such as the recording annotations that track where there is data. The representation of a recording is split into two parts: a segment consisting of a channel id, start time and end time, and a recording annotation tied to the segment, indicating that there is recorded data for the segment. This is the basic architecture of the HSP database: annotations, segments, and the linkage between them, called groundings. TotalRecall has proven indispensable for browsing the recorded data, but it is also a powerful annotation tool. By selecting a portion of audio or video in channel, the user can create a segment, an associated annotation, and save this new metadata to the database. One key annotation is a transcription, which stores the transcript text. In TotalRecall, a user can select a portion of audio containing speech (usually easy to identify in the spectrograms), listen to the audio, create a segment and a transcription annotation, and transcribe the speech. A user may also want to ground a speaker annotation to this segment, labeling who was talking in the segment. In fact, with customizable hotkeys for creating segments and annotations, transcribing speech in this way is faster (Kubat et al., 2007) than with CLAN or Transcriber, two commonly used transcription tools (MacWhinney, 2000; Barras et al., 2001).

In addition to audio annotations such as transcriptions and speaker labels, TotalRecall has also been used for video annotations. In particular, some analyses depend on knowing the child's location in the house. The "where-is-baby" annotations are grounded to segments in video channels that have a clear view of the child and indicate whether he is awake or asleep. Annotation proceeds quickly through a combination of video playback and interpreting video volumes, since once the child has been found in a room only those points where the video volume indicates a room transition need be inspected. Typically, an annotator creates a segment and "where-is-baby" annotation, and then stretches the segment to span the entire duration that the child is present in the video channel. Example "where-is-baby" annotations can be seen in figure 3-1(a).

3.1.2 BlitzScribe

Although TotalRecall was originally used for all annotation, including speech transcription, it was superseded by a more specialized transcription tool called BlitzScribe (Roy and Roy, 2009). BlitzScribe is semi-automatic transcription system that is designed to combine the complementary strengths of human and machine speech processing. While fully automatic speech recognition (ASR) has made impressive performance gains over the years, the quality is still insufficient for speech such as that in the Speechome corpus. Natural, spontaneous speech, in the presence of background noise, collected from microphones at a distance, and with a young child whose early word productions have not all achieved their standard adult form all combine to make the data particularly challenging for automatic speech recognition. Early experiments using off-the-shelf ASR systems resulted in very high word error rates, and although retraining or adapting the models to the data would yield performance gains, this requires a significant amount of training data – namely, speech transcripts. On the other hand, traditional manual approaches to speech transcription are very time consuming, with transcription taking as much as ten to fifty times the actual audio duration (Reidsma et al., 2005; Tomasello and Stahl, 2004; Barras et al., 2001). Of course, faster manual transcription can be performed, as in the case of courtroom stenography, but it is a highly specialized skill and not practical for a long term data annotation project.

User interaction

An analysis of most manual transcription approaches reveals a user interaction model with four basic subtasks. Typically, the transcriber must first FIND speech, often by visually scanning an audio spectrogram or waveform. Using the mouse, the transcriber then SEGMENTs the speech, LISTENs to the audio, and TYPEs the transcript, before moving on to FIND the next speech segment. One bottleneck in this process is physically switching between keyboard and mouse. Another issue is the cognitive load of constantly switching between "input" and "output" subtasks across modalities. This loop is shown in figure 3-2.

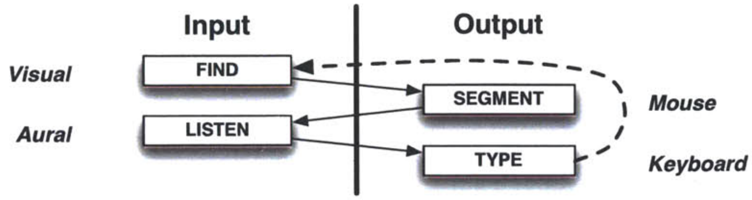
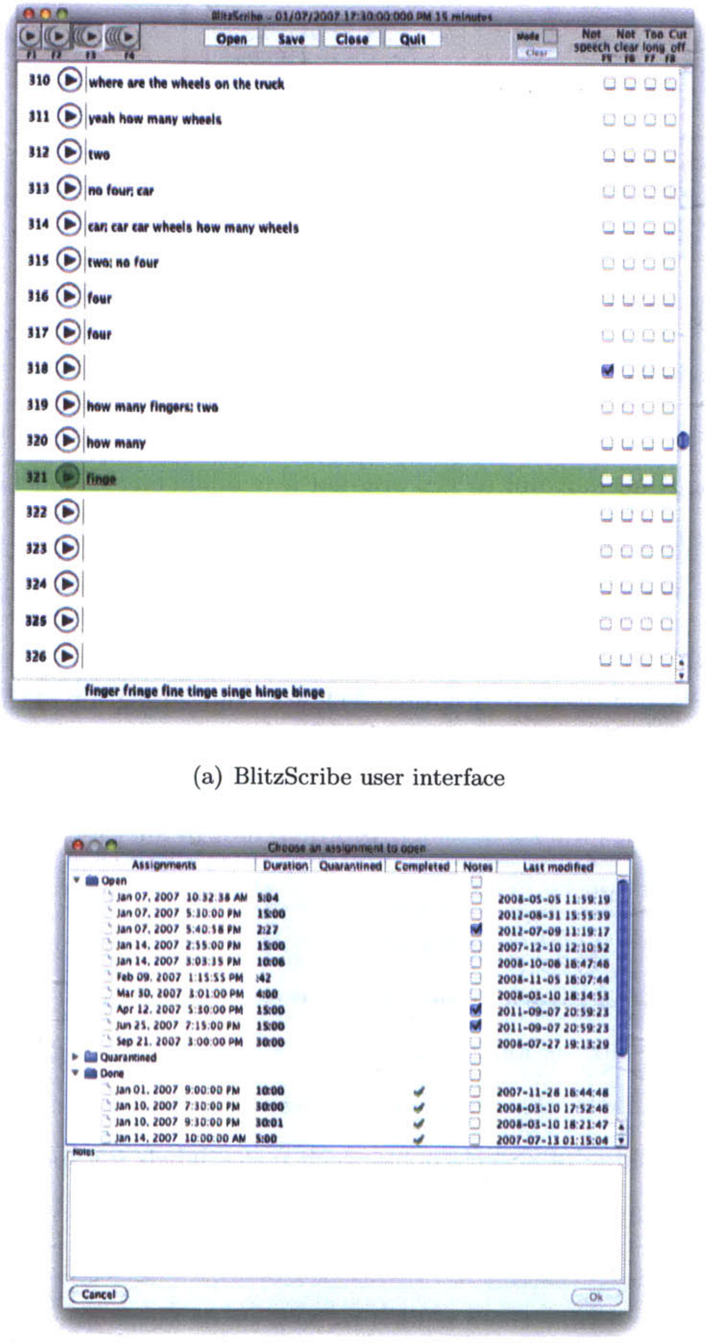


Figure 3-2: The FIND-SEGMENT-LISTEN-TYPE functional decomposition of traditional manual transcription approaches. Repeatedly switching between different input modalities and output devices is inefficient, but automating some of these tasks can speed up the transcription process.

The approach we take in BlitzScribe is to automate those tasks that can be reliably performed by computer, thereby reducing the workload and streamlining the user interface. Specifically, FIND and SEGMENT can be automated, leaving the human to focus on the more appropriate task of interpreting speech and the flow of conversation. BlitzScribe presents the annotator with a sequence of pre-segmented speech clips in a simple user interface, with most of the interaction proceeding via the keyboard. The user plays the current segment by pressing the TAB key, types the transcript, and presses RETURN to advance and automatically begin playing the next segment. Spellchecking is performed on the fly, highlighting and offering suggestions for unknown words. BlitzScribe is shown in figure 3-3.

Speech detection and segmentation

The speech segments presented to the user in BlitzScribe are detected automatically, using a system trained on the Speechome audio to distinguish speech from non-speech. The basic process of speech detection proceeds by first partitioning audio into short, 30 millisecond blocks of samples called "frames", performing feature extraction on each frame to obtain a sequence feature vectors, classifying feature vectors as speech or non-speech, and applying temporal smoothing across labeled frames to obtain a sequence of speech and non-speech segments. The feature vectors consist of standard MFCC features, extracted with code from the Sphinx software package (Walker et al., 2004). In addition to MFCCs, relative power between frequency bands, spectral entropy and zero crossing rate are also used. The frame-level classifier uses a boosted ensemble of decision tree classifiers based on the Weka



	Not speech F5	Not clear F6	Too long F7	Cut off F8
310 where are the wheels on the truck	□	□	□	□
311 yeah how many wheels	□	□	□	□
312 two	□	□	□	□
313 no four; car	□	□	□	□
314 car; car car wheels how many wheels	□	□	□	□
315 two; no four	□	□	□	□
316 four	□	□	□	□
317 four	□	□	□	□
318	☑	□	□	□
319 how many fingers; two	□	□	□	□
320 how many	□	□	□	□
321 finge	□	□	□	□
322	□	□	□	□
323	□	□	□	□
324	□	□	□	□
325	□	□	□	□
326	□	□	□	□

finger fringe fine tinge singe hinge binge

(a) BlitzScribe user interface

Choose an assignment to open

Assignments	Duration	Quarantined	Completed	Notes	Last modified
▼ Open				□	
Jan 07, 2007 10:32:38 AM	5:04			□	2008-05-05 11:59:19
Jan 07, 2007 5:30:00 PM	15:00			□	2012-08-31 15:55:39
Jan 07, 2007 5:40:58 PM	2:27			☑	2012-07-09 11:19:17
Jan 14, 2007 2:55:00 PM	15:00			□	2007-12-10 12:10:52
Jan 14, 2007 3:03:35 PM	10:06			□	2008-10-06 16:47:46
Feb 09, 2007 1:15:55 PM	:42			□	2008-11-05 16:07:44
Mar 30, 2007 3:01:00 PM	4:00			□	2008-03-10 18:34:53
Apr 12, 2007 5:30:00 PM	15:00			☑	2011-09-07 20:59:23
Jun 25, 2007 7:15:00 PM	15:00			☑	2011-09-07 20:59:23
Sep 21, 2007 3:00:00 PM	30:00			□	2008-07-27 19:13:29
▶ Quarantined				□	
▼ Done				□	
Jan 01, 2007 9:00:00 PM	10:00		✓	□	2007-11-28 16:44:48
Jan 10, 2007 7:30:00 PM	30:00		✓	□	2008-03-10 17:52:46
Jan 10, 2007 9:30:00 PM	30:01		✓	□	2008-03-10 18:21:47
Jan 14, 2007 10:00:00 AM	5:00		✓	□	2007-07-13 01:15:04

(b) BlitzScribe assignment chooser

Figure 3-3: The BlitzScribe user interface and assignment chooser. The interface shows segment 321 playing while the transcriber types, with spelling suggestions displayed below. Segment 318 is marked as "not speech".

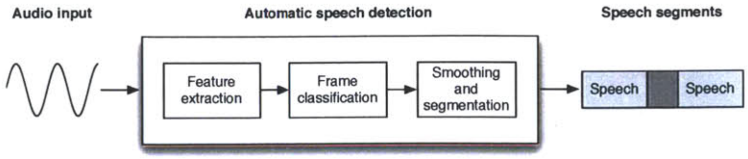


Figure 3-4: Speech detection pipeline. Audio is first partitioned into short frames and converted to feature vectors, each frame's feature vector is then classified, and then the sequence of classifications is smoothed and grouped into speech and non-speech segments.

(Witten and Frank, 2005) Java toolkit. Temporal smoothing is cast as a cost minimization problem, and solved efficiently using a dynamic programming algorithm. The cost structure favors labelings with fewer transitions between speech and non-speech frame labels, but also prefers labelings that are faithful to the original frame labeling produced by the classifier. However, speech segments that are longer than a threshold (5 seconds) are split into shorter segments at points where the speech classification has lower certainty. This is to ensure that, in BlitzScribe, segments are short enough that an annotator usually only need listen to a segment once to fully transcribe it. Figure 3-4 shows a block diagram of the speech detection component.

Note that the Speechome audio is actually recorded from 14 microphones, and given the size and layout of the home speech in one room is generally picked up by multiple microphones. We adopted a simple strategy of dynamically choosing the "best" microphone over time, effectively collapsing the 14-track audio stream to a single stream. The best microphone is determined as the one that is loudest at any given moment, with a penalty for excessive switching between channels. This is accomplished via a cost-minimization scheme similar to the temporal smoothing of speech/non-speech frame labels for speech segmentation. A shortcoming of this simple strategy is in cases where there are multiple, parallel conversations in different parts of the house, since there would be no clear "best" audio channel. Fortunately, at any given time there tended to be a single locus of social interaction, likely due to family size, house layout and child-focused participant behavior. More detail on this process and the speech detection algorithm can be found in (Roy, 2007).

The speech segments that are detected are stored in the central HSP database, using the same data framework as other annotations. Annotations and segments store information about their creator and whether they were produced by a human or machine, so automatically detected speech segments can be differentiated from those identified by a human.

Speech detection errors

One issue with a fully automatic speech detection system is the potential for detection and segmentation errors. In fact, any annotation process, whether manual or automatic, may result in errors and in some cases there is no clear “correct” annotation. We shall return to this when considering speech transcription quality, but the situation is more straightforward for speech detection. A segment detected as speech that does not contain a human vocalization is a false positive error, and human vocalizations that are *not* detected count as false negative errors. In addition, the segment endpoints may be poorly placed, including too much non-speech or cutting off speech. Between false positive and false negative errors, the latter are far more problematic. One key benefit of BlitzScribe is that it eliminates the need for humans to scan vast amounts of audio for speech, and as such it must minimize false negative errors. However, there is often a natural tradeoff between these types of errors, and our system is tuned to minimize false negatives at the expense of producing more false positives. False positives are easily handled in BlitzScribe; they correspond to a segment that does not contain speech, and simply leaving the segment blank and advancing to the next marks the segment as “not speech”. Segments which are cut off can often be repaired by merging adjacent segments, although in our annotation process this is discouraged and instead we rely on transcription conventions for identifying such segments. Segments that are cut off or are too long can be marked in BlitzScribe, potentially to be processed later.

The process of transcribing with BlitzScribe results in an annotation for every detected segment, either associating a transcript or a “not speech” label to the segment. Although the transcript text is the purpose of speech transcription, segments with transcripts or with non-speech labels also constitute valuable human annotated training data for the speech detector. The speech detector can be retrained periodically on more and more training data as transcription proceeds, in turn improving overall system performance. Figure 3-5

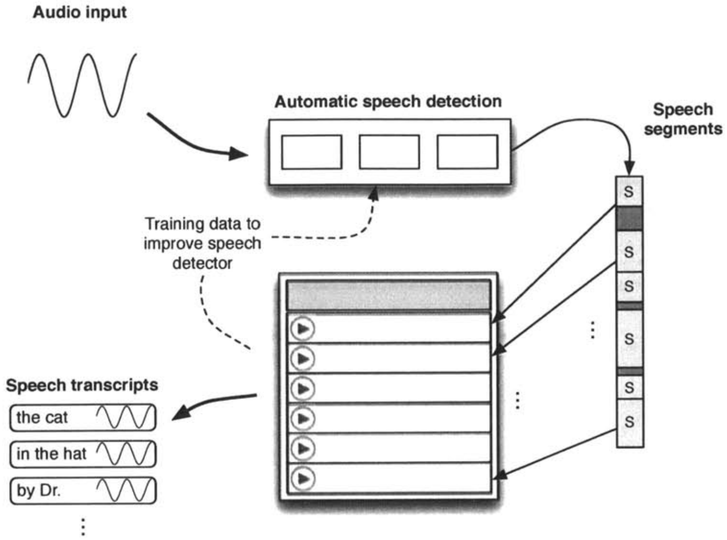


Figure 3-5: BlitzScribe "human-machine collaborative" system overview. The speech detector finds and segments speech from the audio stream and stores the segments into a central database. Speech segments are loaded into the BlitzScribe annotation tool and transcribed, yielding both speech transcripts as well as training data for improving the speech detector.

shows the overall BlitzScribe architecture, described in more detail in (Roy, 2007). We refer to this streamlined combination of human and machine annotation processes as "human-machine collaboration", and it is a general paradigm we have adopted for many aspects of the Human Speechome Project.

Transcription performance

BlitzScribe was built for rapid speech transcription, replacing TotalRecall as our primary transcription tool. Its design was motivated by perceived bottlenecks in traditional transcription approaches, and the expected speed gains were borne out experimentally. We considered two speed measures and two modes of operation, comparing BlitzScribe to CLAN (MacWhinney, 2000) and Transcriber (Barras et al., 2001). The two approaches we considered were "safe" mode, in which those portions of the audio that were *not* detected as speech were manually scanned in TotalRecall for any missed speech, and "fast" mode in which potentially missed speech was not considered. Speed was measured as time spent transcribing relative to either the duration of the audio or the duration of just the speech.

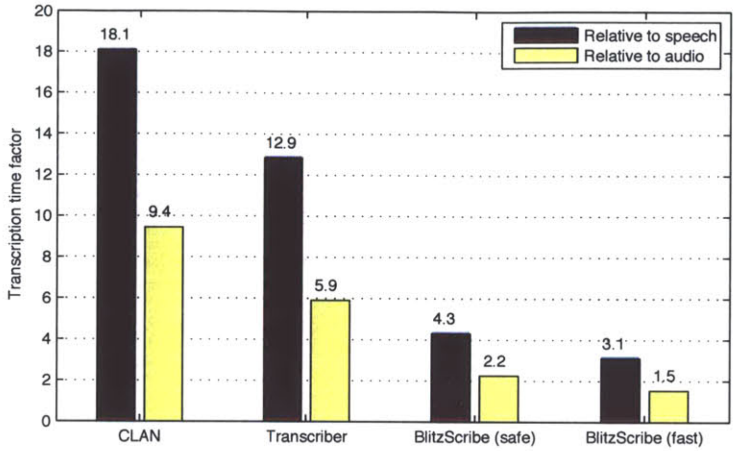


Figure 3-6: Performance of BlitzScribe relative to CLAN and Transcriber. Performance is measured as transcription time relative to the audio duration, or relative to the duration of detected speech. BlitzScribe is also evaluated in two modes: *safe* mode, where non-speech segments are manually checked for speech using TotalRecall, and *fast* mode, where this step is not performed. In practice, we only use fast mode.

In BlitzScribe, the amount of audio presented to the human transcriber depends more on the amount of speech in an assigned block of audio than on the duration of the audio itself. In tools that require scanning the audio manually, both audio duration and speech density contribute to transcription time. Our findings are shown in figure 3-6, with the bottom line result that "fast" mode, our primary means of transcription, is roughly 3–6 times faster than typical approaches. For our data, a high-performing annotator can transcribe 15 minutes of audio in about 30 minutes. Many factors may make this unsustainable, including fatigue, audio assignments with many simultaneous speakers, background noise and so on, but performance at this rate is achievable.

3.1.3 Transcription conventions

When analyzing spoken language, it does not take long to realize that speech is not text. Speech can be coded at multiple levels of detail, from sub-word phonetic annotations to prosody, discourse structure and so on. BlitzScribe is used for orthographic transcription of utterances ranging from roughly .3 – 5 seconds. Segments containing baby babble may not contain words, but we still wish to code these utterances both for analysis and to indicate

`jj`	Use for a word (or phrase) which cannot be understood
`gg`	Use for unknown non-English words
`ff`	Use for a non-speech vocalization, for example, a yawn or cough
`ll`	Use for laughter
;	Use to separate blocks of words by different speakers
-	Use at the beginning or end of a partial word to indicate where it is cut off

Table 3.1: Transcription conventions

that the speech detector did not make an error. Unintelligible speech, overlapping speech by multiple speakers, and non-standard words must be consistently coded. Achieving high quality annotation requires a set of conventions appropriate to the data and transcription style.

Our core conventions are designed to be easy to learn and to support fast transcription. Common occurrences such as non-speech vocalizations and unintelligible speech are given special codes that are easy and fast to type: `ff` and `jj` respectively. Segments with multiple speakers use a ; to separate blocks of words by each speaker, with blocks ordered by the order of occurrence of the first word in each block when the speech overlaps. Sometimes, if the speech segment endpoints are not well placed, the first or last word in the utterance is cut off. In cases where most of the word is present and there is no ambiguity, the word will be transcribed as if it was not cut off. For more severe cases, the cut off part of the word will be replaced with a hyphen and only partially spelled. Thus, if the first word begins with a hyphen or the last word ends with a hyphen, it indicates a cut-off word. Table 3.1 summarizes these core conventions.

The core conventions are frequently used and cover many cases that a transcriber may encounter. But there is far more variety and ambiguity in the data that must be standardized, such as people's names and nicknames, diminutives, machine and animal sounds, special words in other languages, and many other meaningful, consistent, yet non-standard forms of spoken language. Although some of these were known in advance and could be documented and shared with transcribers, many only emerge through exposure to the recordings. With transcribers on the "front lines" and often the first to encounter potential trends in the data, it was important to provide a framework for sharing information. If a transcriber encounters an unclear word they may transcribe it as jj, but if another transcriber has previously recognized this word and shares it, then everyone benefits. For this reason a central, collaborative transcription website was set up in which any annotator can contribute or modify the content. Instead of all questions routing first to the transcription managers, many questions could be answered by other annotators, or ambiguous issues could be collected and then discussed with transcription managers. A Google Site was set up as the central repository for the "how-to" of HSP annotation. In fact, the original BlitzScribe documentation was completely rewritten on the website as a step-by-step guide, with screenshots, by several of the annotators. This site was also used for scheduling time on transcription machines, providing administrative information, sharing relevant publications, and so on. More than 60 transcribers have worked with the Speechome data over the course of several years, and providing a flexible system for sharing information and updating conventions as necessary has been invaluable.

3.1.4 Speaker identification

One crucial aspect of speech transcription that has not been discussed is speaker labeling. For most analyses, it is not just what is said that matters, but who said it. With manual transcription, speaker labeling is often an integral part of the annotation process. For example, text codes for the speakers are part of the transcription in CLAN. This approach could also be used with BlitzScribe with appropriate conventions, however, we were concerned this would slow down the transcription process. Instead, we built a fully automatic speaker ID system to associate a speaker label for each transcribed segment. In addition, a speaker ID labeling tool was built for manual annotation following the BlitzScribe model.

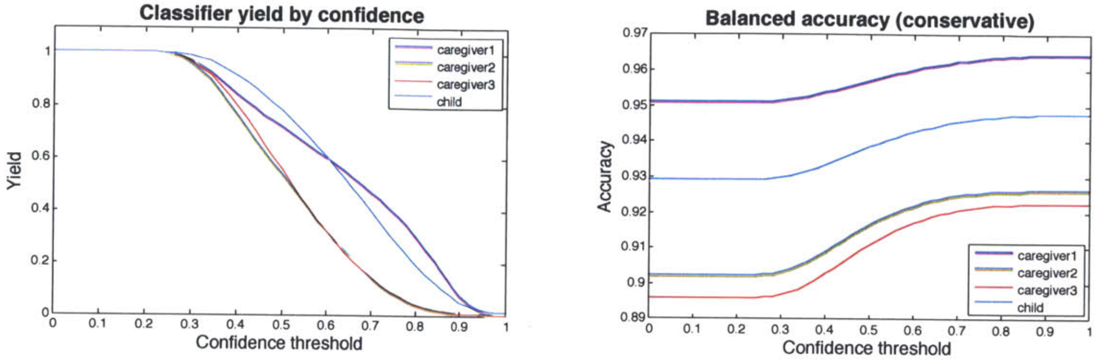
Two versions of the speaker ID system were built, the first in 2007 by myself and a second, improved version in 2009 by Matt Miller. These systems are similar to the speech detector, but since the speech segments are already provided segmentation is unnecessary. As with speech detection, MFCCs are used as input features, along with delta and delta-delta MFCCs. The speaker ID training set consists of speech segments with manually annotated speaker ID labels. These audio segments are split into a sequence of 30 ms frames with 15 ms overlap, and each frame is converted to a feature vector with the segment speaker label.

In the original speaker ID system, a boosted ensemble of decision trees were trained on the four primary speakers. The decision trees labeled each frame and provided a confidence value; the label that had the maximum accumulated confidence over all frames was assigned to the segment. The fraction of this maximum accumulated confidence out of the total confidence for all frames was used as the segment confidence.

Although this system was a useful starting point, significant improvements were made by Matt Miller in 2009 using a frame-level classifier based on Gaussian mixture models (GMMs). This system begins by training a single GMM on all data, often called a Universal Background Model (UBM), then uses this model as the starting point to train speaker-specific GMMs. One advantage of a generative probabilistic model such as a GMM is that it may smooth over noise in the training data. It also provides a principled way of combining frames into the final segment label in terms of the overall likelihood function for each class.

Automatic speaker ID performance

To assess the speaker ID system, we consider several performance measurements. When the classifier is invoked on a segment, it returns both a classification and a confidence value. In general, high confidence classifications are more accurate, and for some analyses we wish to use only higher quality annotations. A confidence threshold can be used to filter out potentially low accuracy speaker classifications. On the other hand, raising this threshold may discard too much data. Figure 3-7(a) shows the *yield* as a function of confidence threshold. Yield is the fraction of classifications for a target speaker that are above threshold. The maximum confidence value is 1, thus the yield drops to 0 as the confidence threshold increases. Note that the yield does not capture actual classifier performance – classifications for a target may be above threshold but also incorrect. Figure 3-7(b) shows one measure of accuracy as a function of confidence. This measure, called *balanced accuracy*, equally balances the fraction of correct positive guesses and the fraction of correct negative guesses



(a) Yield, showing the fraction of classifications above confidence threshold as the threshold is varied.

(b) Balanced accuracy, showing the equally weighted fraction of correctly guessed positives and negatives as the confidence threshold is varied.

Figure 3-7: Speaker ID yield and accuracy as a function of confidence threshold.

by the classifier. Balanced accuracy is useful because for a given target speaker, segments for the other speakers constitute negative examples resulting in many more negative than positive examples. Balanced accuracy corrects for this issue. Here we are reporting a "conservative" measure of balanced accuracy, in which the classifier is not penalized for withholding a guess if the classification is below the confidence threshold. Accuracy is only assessed for the subset of segments where the classification was above threshold. For a given confidence threshold, the classifier yield and the corresponding accuracy give a sense of the system performance.

Another useful measure of classifier performance is an ROC curve, shown in figure 3-8. An ROC curve shows the relationship between the classifier true positive rate (TPr) and false positive rate (FPr). The true positive rate reflects the fraction of all positive examples for a class that are correctly guessed by the classifier, also known as *recall.* The false positive rate is the fraction of negative examples for a class that are incorrectly guessed as the class. For a particular class, always guessing the class will yield TPr = 1 but also FPr = 1, since all positive *and* negative examples are labeled as positive. On the other hand, never guessing the class yields TPr = 0 and FPr = 0. Most classifiers fall in between these two extremes, with their performance often varying as a function of a threshold (eg. the confidence threshold.) In our case, the classifier is not binary but a multiclass classifier, and the performance does not vary over the full ROC space, hence the title "partial ROC

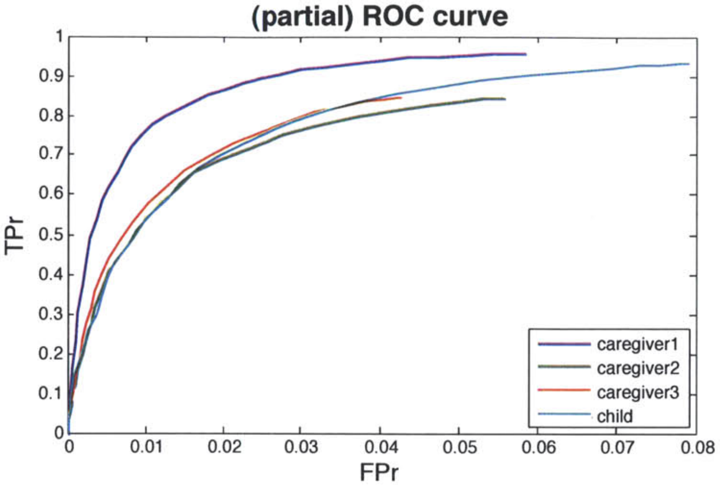
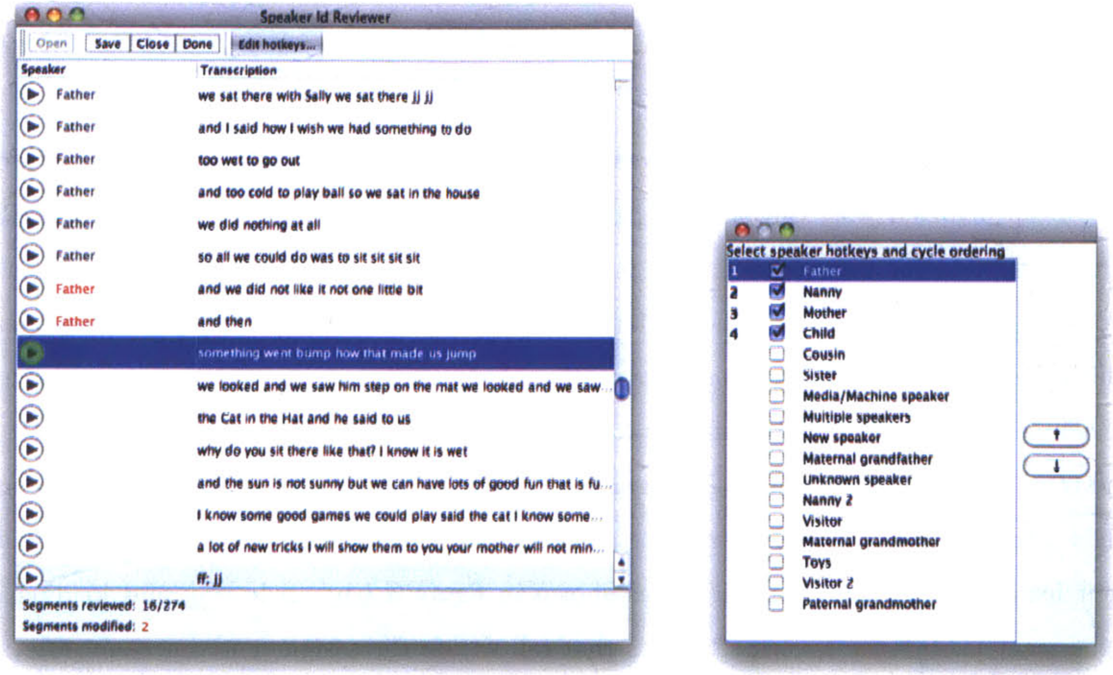


Figure 3-8: Speaker ID ROC curve, showing the relation between the classifier TPr and FPr for various confidence threshold settings.

curve" in the figure.

Manual speaker ID annotation

Although there are only a small number of speakers, obtaining high-accuracy speaker labeling has been surprisingly difficult. To augment the fully automatic system, a speaker labeling tool was built. This system fetches speech segments from the database and presents them in a view that is similar to BlitzScribe and is operated in a similar fashion. Speaker labels already annotated by the current user are displayed in dark gray. In an earlier version of the tool, machine generated speaker labels and corresponding confidence values were displayed. This enabled users to sort the segments to focus on low confidence annotations or segments by a particular speaker, but this complicated the user interface and did not seem to improve annotation speed or quality. In this earlier version, we also experimented with default annotations, whereby listening to a segment without changing the machine annotation marked it as correct. However, this behavior would often have unintended consequences and was replaced by a simpler, direct annotation process. The direct annotation process supports three means of entering speaker labels. Speaker labels can be bound to user-customizable hotkeys, the arrow keys can be used to cycle through speakers, or the



Speaker	Transcription
Father	we sat there with Sally we sat there jj jj
Father	and I said how I wish we had something to do
Father	too wet to go out
Father	and too cold to play ball so we sat in the house
Father	we did nothing at all
Father	so all we could do was to sit sit sit sit
Father	and we did not like it not one little bit
Father	and then
	something went bump how that made us jump
	we looked and we saw him step on the mat we looked and we saw...
	the Cat in the Hat and he said to us
	why do you sit there like that? I know it is wet
	and the sun is not sunny but we can have lots of good fun that is fu...
	I know some good games we could play said the cat I know some...
	a lot of new tricks I will show them to you your mother will not min...
	ff; jj

Segments reviewed: 16/274
Segments modified: 2

Select speaker hotkeys and cycle ordering

1	☑	Father
2	☑	Nanny
3	☑	Mother
4	☑	Child
	□	Cousin
	□	Sister
	□	Media/Machine speaker
	□	Multiple speakers
	□	New speaker
	□	Maternal grandfather
	□	Unknown speaker
	□	Nanny 2
	□	Visitor
	□	Maternal grandmother
	□	Toys
	□	Visitor 2
	□	Paternal grandmother

Figure 3-9: The Speaker ID annotation tool. The primary window on the left shows the current speaker ID annotations in dark gray and the newly entered annotations in red. Transcripts containing multiple speakers, such as the last one displayed, are grayed out and cannot be annotated. The window on the right allows configuring the keyboard hotkeys, whether speakers should be in the cycle, and the cycle order.

mouse can be used to choose the speaker from a list. Annotation proceeds quickly, and while we did not undertake an analysis of annotation speed, the tool has been an effective supplement for the automatic system. Figure 3-9 shows the tool user interface.

3.2 Transcription management

Just as collecting the data for the Human Speechome Project has been a large-scale project with significant new infrastructure development, annotating the Speechome data has also required extensive work developing new tools and methods. The centerpiece of our annotation work is BlitzScribe, but an entire ecosystem of tools and practices has grown up around BlitzScribe to support annotation.

3.2.1 Assignments in BlitzScribe

Since transcription with BlitzScribe was launched in June 2007, more than 60 annotators have been involved with the project, primarily working as transcribers using BlitzScribe. A clean system for distributing work to transcribers and tracking individual and overall progress has proven essential. The fundamental unit of work in our system is the *assignment*, which is a block of time requiring transcription that is assigned to a particular transcriber. Assignments are typically 15 minutes long, although we originally experimented with 30 and 60 minute long assignments. Assignments are stored in the central HSP database, and when a transcriber launches BlitzScribe and logs in with their unique account information, their list of assignments is retrieved as shown in figure 3-3(b). Selecting and loading an assignment fetches the speech segments that fall within the assignment timespan, filtering out those speech segments the child wouldn't be able to hear. Only speech occurring in channels near to the child, and when the child is awake, are presented to the transcriber. This is accomplished by filtering speech segments against the "where-is-baby" annotations.

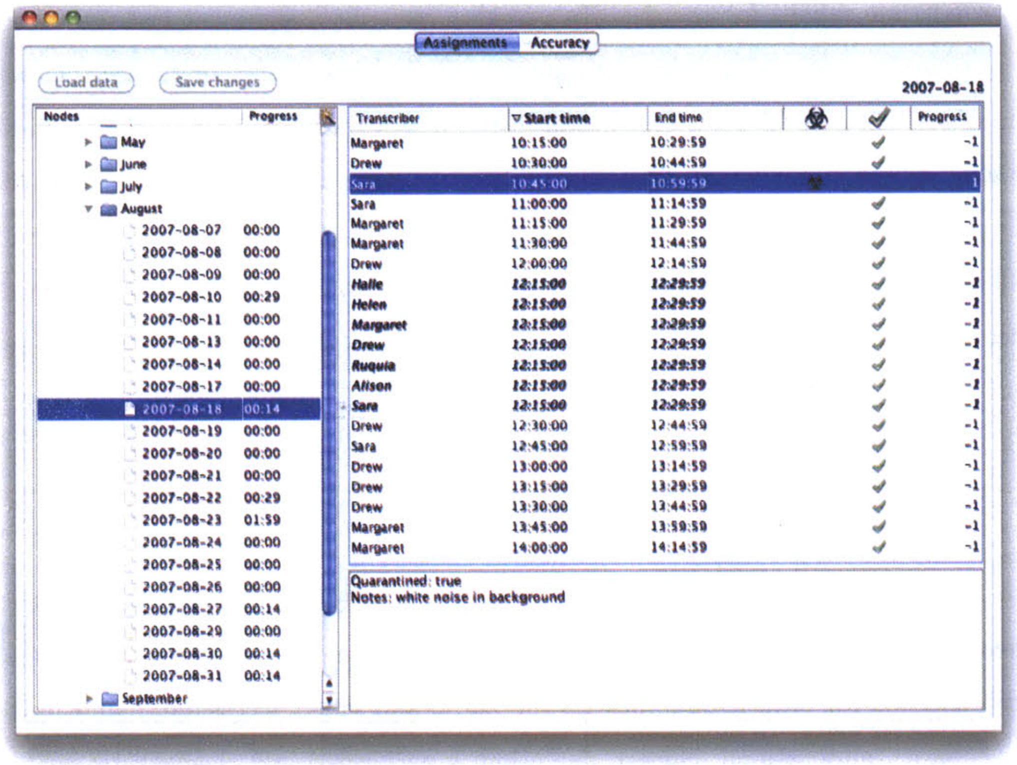
Once segments are loaded into BlitzScribe, transcription proceeds as described above. Transcribers can save their work to the database for later retrieval, and when they have transcribed all speech in an assignment it can be marked as complete and will be removed from their list of open assignments. Notes can also be associated with an assignment, with annotators often commenting on the audio or speech detection quality, whether there is interesting or sensitive content, or anything else they wish to record. For some assignments, annotators may choose to "quarantine" rather than complete the assignment. Assignments with very poor speech detection, significant background noise, or other issues that may result in poor transcription quality can be quarantined. One benefit of the quarantine functionality is that it reduces the burden on the transcription manager to inspect speech detection quality before creating and distributing assignments. Instead, annotators can assess the assignment quality and if they choose to quarantine it they can document their reasons in the assignment notes.

3.2.2 Assignment creation and distribution

Assignment creation and distribution is primarily a manual process performed by transcription managers. The basic unit of data to be processed is a day of recording, and the goal of this process is to obtain a full set of detected speech segments for the day, to verify that "where-is-baby" annotations are complete, and to produce a full set of assignments covering all recorded time ranges for the day. In practice, we split assignment creation across two roles: a content management role, performed by Halle Ritter, a senior transcriber who had been involved with the project for several years, and a technical manager role, performed by myself. The content manager uses TotalRecall to scan candidate days to check that data is present, that "where-is-baby" annotations are complete and to update them if not, and then to update a shared spreadsheet with days that are ready to be processed. The technical manager fetches days from the spreadsheet, runs the speech detection system and imports the detected speech segments into the database, and uses TotalRecall to check quality and create assignments, putting them in the "assignment inbox." The status of the day is updated to "assigned" in the spreadsheet.

An important concept that has made this process work is a clear sequence of stages each day must pass through, with the final stage being "assigned". Before a day is assigned, it is first checked for data availability (is the audio and video accessible?) and up to date "where-is-baby" annotations. Speech detection is performed and checked for quality, and finally assignments are created and put in the assignment inbox. If there are major problems the entire day may be withheld, or particular assignments may be left unassigned.

Once assignments are in the inbox, they can be distributed to annotators by a transcription manager. This task was primarily performed by Karina Lundahl, who oversaw the transcription team. Using the "HSP Reporter" tool, assignments are moved out of the inbox to particular transcribers. Assignments can also be duplicated to multiple transcribers for assessing inter-annotator agreement. The tool provides a view of the open, quarantined and closed assignments for a particular annotator or for a particular day. A screenshot of the HSP Reporter is shown in figure 3-10.



Nodes	Progress
▶ May	
▶ June	
▶ July	
▼ August	
2007-08-07	00:00
2007-08-08	00:00
2007-08-09	00:00
2007-08-10	00:29
2007-08-11	00:00
2007-08-13	00:00
2007-08-14	00:00
2007-08-17	00:00
2007-08-18	00:14
2007-08-19	00:00
2007-08-20	00:00
2007-08-21	00:00
2007-08-22	00:29
2007-08-23	01:59
2007-08-24	00:00
2007-08-25	00:00
2007-08-26	00:00
2007-08-27	00:14
2007-08-29	00:00
2007-08-30	00:14
2007-08-31	00:14
▶ September	

2007-08-18

Transcriber	▽ Start time	End time	☣	✓	Progress
Margaret	10:15:00	10:29:59		✓	-1
Drew	10:30:00	10:44:59		✓	-1
Sara	10:45:00	10:59:59	☣		1
Sara	11:00:00	11:14:59		✓	-1
Margaret	11:15:00	11:29:59		✓	-1
Margaret	11:30:00	11:44:59		✓	-1
Drew	12:00:00	12:14:59		✓	-1
Halle	***12:15:00***	***12:29:59***		✓	***-1***
Helen	***12:15:00***	***12:29:59***		✓	***-1***
Margaret	***12:15:00***	***12:29:59***		✓	***-1***
Drew	***12:15:00***	***12:29:59***		✓	***-1***
Ruquia	***12:15:00***	***12:29:59***		✓	***-1***
Alison	***12:15:00***	***12:29:59***		✓	***-1***
Sara	***12:15:00***	***12:29:59***		✓	***-1***
Drew	12:30:00	12:44:59		✓	-1
Sara	12:45:00	12:59:59		✓	-1
Drew	13:00:00	13:14:59		✓	-1
Drew	13:15:00	13:29:59		✓	-1
Drew	13:30:00	13:44:59		✓	-1
Margaret	13:45:00	13:59:59		✓	-1
Margaret	14:00:00	14:14:59		✓	-1

Quarantined: true
Notes: white noise in background

Figure 3-10: The HSP Reporter, showing assignments for August 18, 2007. The currently selected assignment has been quarantined. Completed assignments are checked. Bold, italic assignments have been duplicated to multiple annotators for calculating inter-annotator agreement.

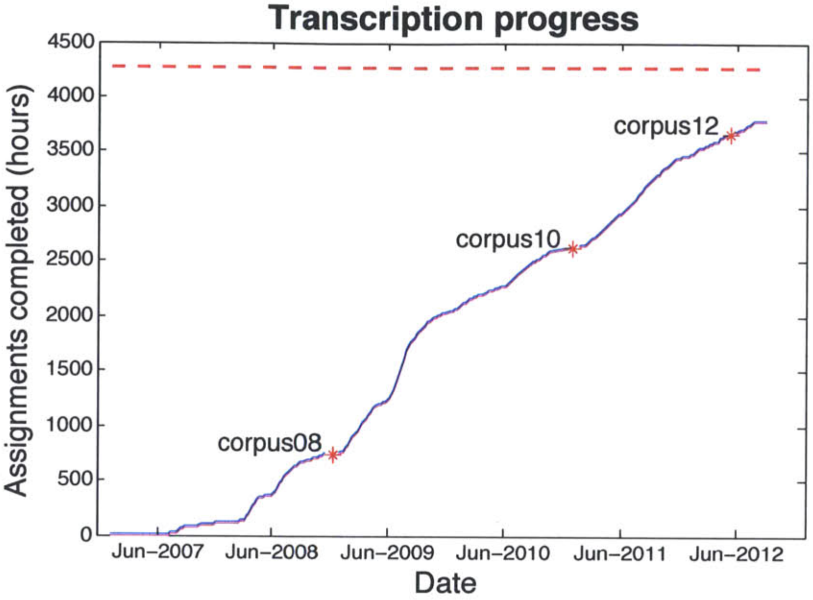


Figure 3-11: Transcription progress over time, from 2007 to 2012. The dashed red line indicates the total number of hours of recorded (multichannel) audio in the child's 9–24 month age range. Three snapshots of the corpus were taken at points marked on the graph. As of May 14, 2012 (`corpus12`), transcription of this period was 86% complete.

3.2.3 Monitoring transcription progress and quality

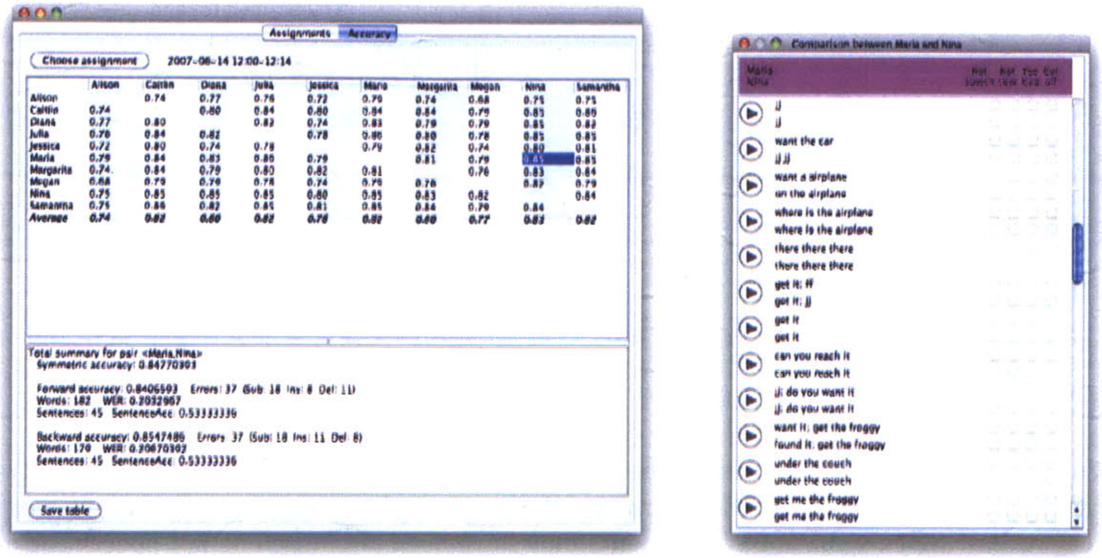
Transcribing the Speechome data is a long-term project, and tracking both progress and quality has been an important element of our methodology. Our primary measure of progress is in terms of closed assignments; marking an assignment as "done" implies that all detected speech segments in the assignment time range have been transcribed or marked as not speech. The duration of all closed assignments relative to the total duration of all recorded (multi-channel) audio is the percentage complete. This overall quantity can be broken down by month of recording, which is automatically calculated and emailed weekly. Note that while a single block of time may be assigned to multiple annotators, only one assignment need be marked as complete for the block of time to be considered fully annotated. As of May 14, 2012 transcription of the child's 9–24 month age range is 86% complete, with the progress by month shown in figure 3-11.

In addition to overall progress, transcriber performance is also reported weekly. Transcribers receive a summary of their annotation output, with transcription managers receiving a full report. This includes information for the week including the amount of time spent using BlitzScribe, the number of assignments closed, and the number of words transcribed. In addition, the cumulative number of words transcribed is also reported. These weekly reports serve the dual goal of helping transcription managers monitor progress as well as to motivate annotators. For annotators, we hope this builds accountability into the process and provides some perspective on the overall annotation enterprise.

Our primary measure of transcription quality is inter-transcriber agreement. Inter-transcriber agreement measures the similarity between transcripts by two different annotators for the same audio. A standard way of comparing two strings is using the Levenshtein distance (or edit distance) over either characters or words. This distance captures the minimum number (or minimum cost) set of "edits" required to align one string with another. For our purposes, the units are case insensitive words and the edit operations are word insertion, deletion or substitution, each with an associated cost. The Levenshtein distance between two identical strings is 0.

In automatic speech recognition, system performance is often measured relative to a "gold standard" set of human produced transcripts. The NIST `sclite` tool (Fiscus, 2007) can be used to align hypothesis transcripts against corresponding reference transcripts. In our case, there is no gold standard reference transcript, so we treat each as the hypothesis and reference in turn, calculating the alignment twice and averaging the resultant scores to obtain a symmetric distance. However, rather than reporting and averaging the edit distance values, we use the word accuracy, which is the fraction of words in the reference transcript that were correctly placed in the hypothesis transcript. This quantity is easier to interpret. Of course, not all errors are of equal significance. Certain transcription codes indicate vocalizations such as laughter, non-English speech and so on, and these can reasonably differ without impacting quality. Before calculating the alignment score the transcripts are canonicalized. Thanks to a Java implementation of the NIST alignment algorithm in Sphinx (Walker et al., 2004), this functionality is built into the HSP Reporter.

Using the HSP Reporter, assignments are periodically duplicated to multiple transcribers. In particular, when a new transcriber is hired their first assignments are usually duplicates



2007-06-14 12:00-12:14

	Alison	Caitlin	Diana	Julia	Jessica	Maria	Margarita	Megan	Nina	Samantha
Alison		0.74	0.77	0.76	0.72	0.79	0.74	0.68	0.75	0.75
Caitlin	0.74		0.80	0.84	0.80	0.84	0.84	0.79	0.85	0.86
Diana	0.77	0.80		0.82	0.74	0.83	0.79	0.79	0.85	0.82
Julia	0.76	0.84	0.82		0.78	0.86	0.80	0.78	0.85	0.85
Jessica	0.72	0.80	0.74	0.78		0.79	0.82	0.74	0.80	0.81
Maria	0.79	0.84	0.83	0.86	0.79		0.81	0.79	0.85	0.85
Margarita	0.74	0.84	0.79	0.80	0.82	0.81		0.76	0.83	0.84
Megan	0.68	0.79	0.79	0.78	0.74	0.79	0.76		0.82	0.79
Nina	0.75	0.85	0.85	0.85	0.80	0.85	0.83	0.82		0.84
Samantha	0.75	0.86	0.82	0.85	0.81	0.85	0.84	0.79	0.84	
Average	*0.74*	*0.82*	*0.80*	*0.82*	*0.78*	*0.82*	*0.80*	*0.77*	*0.83*	*0.82*

```
Total summary for pair <Maria,Nina>
  Symmetric accuracy: 0.84770303

  Forward accuracy: 0.8406593   Errors: 37 (Sub: 18 Ins: 8 Del: 11)
  Words: 182   WER: 0.2032967
  Sentences: 45   SentenceAcc: 0.53333336

  Backward accuracy: 0.8547486   Errors: 37 (Sub: 18 Ins: 11 Del: 8)
  Words: 179   WER: 0.20670392
  Sentences: 45   SentenceAcc: 0.53333336
```

(a) Inter-transcriber agreement matrix.

Comparison between Maria and Nina

Maria
Nina

jj
jj

want the car
jj jj

want a airplane
on the airplane

where is the airplane
where is the airplane

there there there
there there there

get it; ff
got it; jj

got it
get it

can you reach it
can you reach it

jj; do you want it
jj; do you want it

want it; get the froggy
found it; get the froggy

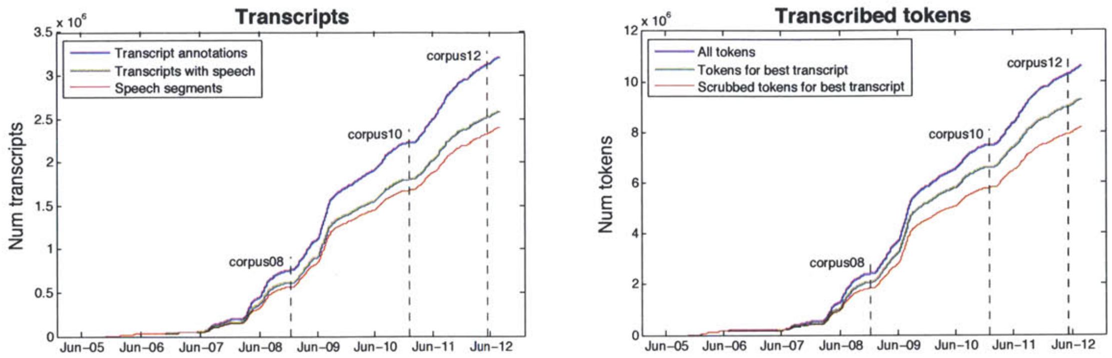
under the couch
under the couch

get me the froggy
get me the froggy

(b) Comparing transcripts from two annotators for the same speech segments.

Figure 3-12: Assessing transcription quality and transcriber performance in the HSP Reporter.

of assignments already completed by experienced transcribers. Inter-transcriber agreement can then be calculated and displayed as a pairwise distance matrix in the tool, as shown in figure 3-12(a). The average agreement for a transcriber is shown in the bottom row, which is a quick way of assessing that transcriber's agreement with the rest of the annotators. A low overall agreement score may indicate a misunderstanding of transcription conventions or some other issue – but whatever the case, it points out work that should be checked by a transcription manager. Right-clicking on a cell in the agreement matrix shows the transcriptions by each transcriber, along with a button for playing back the audio, shown in figure 3-12(b). Transcription managers can use this feature to compare the actual transcriptions. Another useful interpretation of the inter-transcriber agreement quantities is as a reflection of overall assignment difficulty. Some assignments have a high overall average agreement while others are lower. In particular, assignments with many simultaneous speakers are often more challenging.



(a) Number of transcript annotations over time

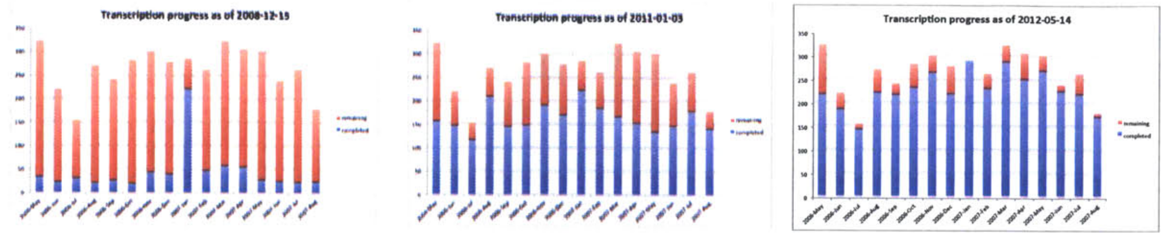
(b) Number of transcribed tokens over time

Figure 3-13: The amount of transcribed speech in the central Speechome database. Transcribed speech can be measured in terms of the number of transcripts or the number of tokens (words). Since some speech segments may have been annotated as not-speech, or may have several transcripts from different annotators, we also show the number of transcripts that actually contain speech and the number of transcribed segments. The total number of tokens, the total for the selected transcriptions, and the total number of scrubbed tokens after removing non-word vocalizations is shown on the right.

3.3 The Speechome Corpus

Soon after recording began, we started to annotate limited portions of the most recently recorded data. These annotations were stored in a central database, from the beginning of the project to the present. The bulk of our annotation work has been in transcribing the speech; figure 3-13 shows the number of transcribed segments and total tokens in the central database since we began in 2005.

While the central database spans the entire 3 year recording period, the focus of our annotation work has been on the subset of the data spanning the child's 9–24 month age range. This annotated subset, from May 1, 2006 through August 31, 2007, captures the child's first words up to multiword speech. We refer to this annotated subset as the *Speechome Corpus*. The Speechome Corpus spans 488 days, of which data was recorded for 444 days – about a 91% rate of recording. The audio, treated as a single 14 track recording, is 4260 hours long and averages to about 9.6 hours of audio recorded per day. The months that comprise the Speechome Corpus are highlighted in figure 2-6.



(a) Progress as of Dec 15, 2008 for `corpus08`. 17% complete.

(b) Progress as of Jan 3, 2011 for `corpus10`. 62% complete.

(c) Progress as of May 14, 2012 for `corpus12`. 86% complete.

Figure 3-14: Transcription progress in terms of hours of assignments closed relative to audio recorded in the Speechome Corpus, for `corpus08`, `corpus10` and `corpus12`.

Annotation has been an ongoing process, but at various points we have taken "snapshots" of the data for our analyses. Three corpora, dubbed `corpus08`, `corpus10`, and `corpus12`, were prepared in December 2008, December 2010, and May 2012 respectively. One reason for taking a snapshot of the central database is for convenience: fixing the data simplifies the analysis and ensures that results can be reproduced at a later date. Furthermore, the annotated data in the corpus still requires some preparation before it can be used. A key step in preparing data is selecting the annotation to use for a segment when there are multiple annotations of the same type. For example, some speech segments have multiple transcriptions (usually from assignments distributed to multiple transcribers), and the "best" one should be selected. This is done programmatically by preferring transcripts that have fewer `ff` and `jj` tokens and more legitimate words. It could also be done by preferring transcripts by certain annotators. Similar logic is applied for selecting the speaker ID annotation – a human generated annotation is preferred over a machine annotation, and among machine annotations the most recent one is preferred, since a new speaker ID system is only deployed when its performance is improved.

One summary characteristic of these three corpora is the transcription progress at the time of each corpus snapshot. In figure 3-11 and 3-13 we saw transcription progress over time in terms of closed assignments, transcripts, and tokens, with markers for `corpus08`, `corpus10` and `corpus12`. However, a more detailed picture of just these three snapshots is provided by figure 3-14, which indicates the hours of completed assignments relative to hours of audio recorded for each month in the Speechome Corpus.

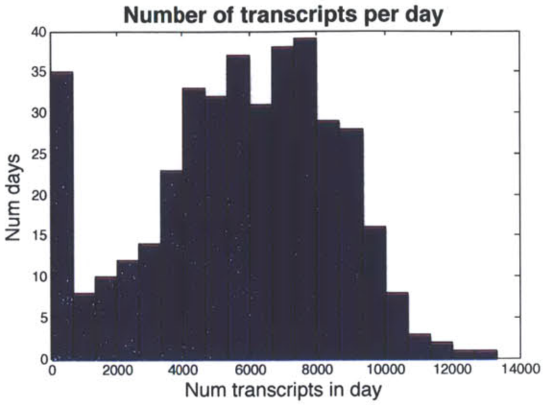


Figure 3-15: Histogram of the number of transcripts per day for `corpus12`. The peak in the first bin is due to partially transcribed days.

3.3.1 The content of `corpus12`

The data used for this thesis is the most recent corpus snapshot, `corpus12`, which covers 86% of the recorded 9–24 months of audio as measured by closed assignments. This corpus contains transcripts for 400 days, although in general not all of these days were fully transcribed. Figure 3-15 shows a histogram of the number of transcripts per day for the 400 days, with a clear peak in the first bin for days with only a small number of utterances. This is a result of some days being only partially transcribed at the time of export[1].

This corpus contains a total of 2,279,686 transcripts and 8,759,587 tokens, and these are distributed over the 9–24 months as shown in figure 3-16. These distributions generally follows the distribution of recorded data. For many analyses, some word types are excluded, such as those indicating partial words (beginning or ending with a hyphen) or non-speech vocalizations. Removing these reduces the number of tokens by about 15%, and this "filtered" set consists of 7,411,973 tokens but has roughly the same monthly distribution of token counts. Since removing tokens may leave some transcripts empty, this also reduces the number of transcripts to 1,725,342, a reduction of about 24% relative to the unfiltered set.

[1] In fact, some days do not have *any* completed assignments but still have transcripts, possibly due to transcribers still working at the time of export or older transcripts entered prior to the use of BlitzScribe and assignments.

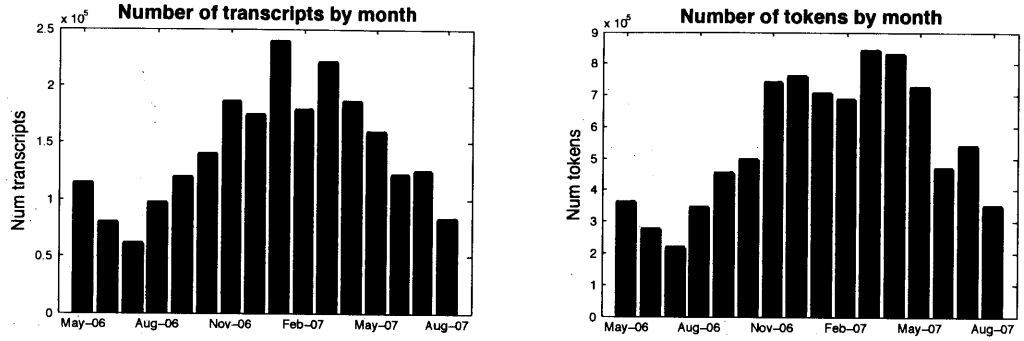


Figure 3-16: Distribution of transcripts and tokens by month in `corpus12`.

Another summary characteristic of `corpus12` is the distribution of transcripts and tokens by speaker. We group the transcripts into 6 groups: the 4 primary speakers, a group for multispeaker utterances, and "other" for any single-speaker transcripts with speaker confidence below .4 or not fitting into the other groups. Using the filtered set described above, figure 3-17 shows the proportion of filtered transcripts attributed to each of the four primary speakers by month. Interestingly, the child's proportion of transcripts increases to about 40% by 24 months. However, this does not mean that the child is responsible for 40% of the *tokens* in later months, since his average number of word tokens per transcript is lower than that of adults. In fact, the child's word token proportion increases steadily over time, but only by August 2007 does it reach an equal share of about 25% of the word tokens. Of course, since the speaker ID system is not perfect it will affect speaker proportions somewhat.

The other two speaker groups – multispeaker and "other" – comprise about 32% and 15% of the total number of transcripts. These proportions do not vary much with time, although the percentage of multispeaker utterances likely increases in months when extended family is visiting and there are more people in the house. Multispeaker utterances are not an issue when considering overall word frequencies, but do pose problem in linking a speaker to the words they said. For such analyses, we generally discard these utterances. In the future, a second pass of annotating the individual speakers for multispeaker utterances may be worthwhile.

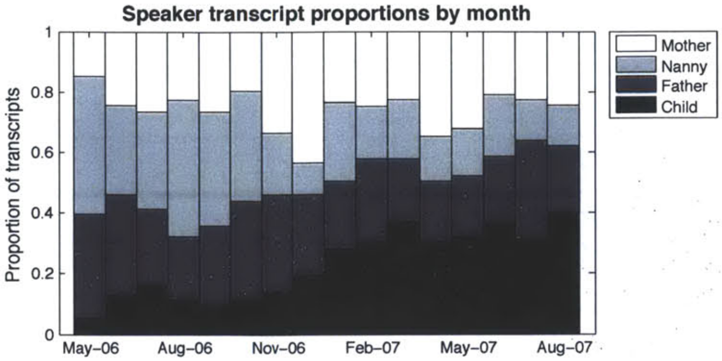
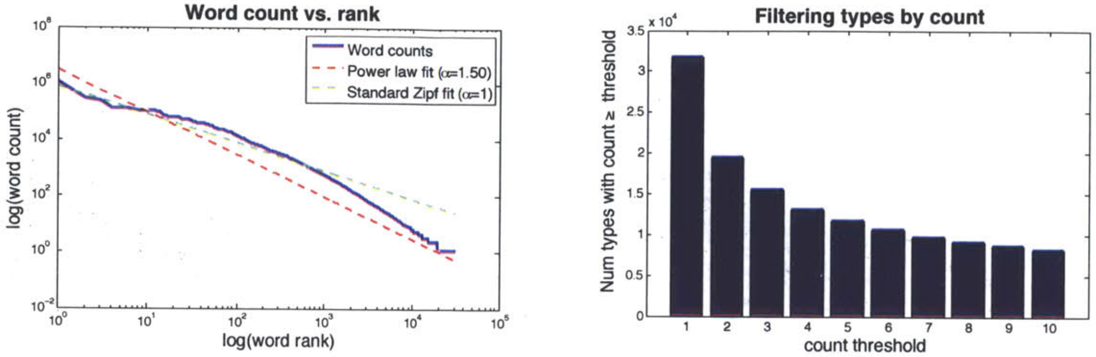


Figure 3-17: Stacked bar graph showing the proportion of transcripts attributed to each speaker by month. The transcripts considered are filtered to remove transcripts consisting only of partial words or non-speech vocalizations.

There are a total of 31,750 word types in the corpus, with the top 10 words making up more than 25% of all tokens, including words like "jj", "you", "the", "I", and "it". The word frequency distribution relative to rank is shown in figure 3-18(a) on a log-log plot since there are relatively few frequent words and many infrequent words. The apparent power law relationship of word i's frequency f_i to its rank r_i was observed by (Zipf, 1949) and dubbed "Zipf's Law". Aligned with the empirical data we show $f_i \propto 1/r_i^{\alpha}$, setting $\alpha = 1$ for the standard Zipf's Law form. However, using the techniques described in (Clauset et al., 2009) to find the parameters for a power law, an alternative fit with $\alpha = 1.5$ is also shown. There has been considerable interest in Zipf's Law and more general power-law relationships in natural phenomena. (Li, 1992) discusses the conditions that can give rise to such relationships in text, and (Clauset et al., 2009) describe methods for fitting and testing these models.

For our purposes, we do not wish to argue that word frequencies in the Speechome data follow a power law. However, such distributions do have "heavy tails", and this is also the case of our word frequency data. There are many word types that occur only once or a few times in the corpus of 7.4 million tokens. Many misspelled words are singletons, since in general there are many ways to misspell a word and often these words are simply mistyped.[2] Of course, many singletons are legitimate but rare words. Nevertheless, many statistical

[2] Interestingly, some of the misspellings that occur more than once are unique to a particular annotator, and are a kind of "signature" misspelling.



(a) Relationship between word frequency and word rank on a log-log plot, along with power laws with $\alpha = 1$ (Zipf's Law) and $\alpha = 1.5$.

(b) The number of word types with a minimum threshold of occurrences, as the threshold is increased.

Figure 3-18: Summary plots for the word types in the Speechome Corpus.

analyses require a minimum number of occurrences for a word, and filtering words by token count quickly reduces the total number of word types, as shown in figure 3-18. This graph shows that removing all singletons reduces the number of unique word types by nearly 40%, to 19,431 words.

3.4 Conclusion

The Human Speechome Project began with the goal of studying language acquisition in the context where it occurs: a child's home. Many of the early challenges have been technical, from building a robust and unobtrusive recording system to designing and managing a large-scale transcription process. The resultant Speechome Corpus is unique in the density and coverage of the audio and video recordings as well as the extent of the annotated data, easily the largest annotated corpus of one child's development to date.

The Speechome Corpus is the foundation for our study, but the real story of early word learning begins with the child and his first words. What words does he learn, when does he learn them, and can we identify factors that help explain "why"?

Chapter 4

The Child's Productive Vocabulary

A child's first words are a clear achievement, and mark an important developmental milestone. But in a sense, word learning is a network of linked processes that do not necessarily proceed in lockstep on their developmental path. Children's earliest word productions may not have reached their adult phonological form as the articulatory system develops, but a caregiver may still consider such productions as "words" if they are used consistently and understandably. Consistent word *meaning* is clearly an important attribute of a word, but expecting absolute alignment between the child's and adult's meaning is an onerous requirement. Indeed, Wittgenstein (1965; 2009) points out the problems in defining a word's meaning, instead arguing that the meaning of a word is its *use*. In this sense, language – and words – acquire meaning as part of the social and cultural context. From a more technical standpoint, the fundamental meaningful unit in a language *isn't* the word but the morpheme, a word's more basic constituents.

Nevertheless, a word is a locus of meaning, and at least in speech, a fundamental constituent of language that stands on its own. As the point where sound and meaning come into contact, a child's mastery of words – both production and comprehension – is their jumping off point into language. But how do children's vocabularies grow over time, and are there any clear patterns of development?

4.1 Early vocabulary growth

Children typically begin producing their first words around their first birthday, and are combining words together near the end of their second year of life. First words and the onset of combinatorial speech are clearly important milestones, but what happens in between? In this section we review selected work on this important period of development.

4.1.1 Vocabulary size and growth rate

One way of characterizing children's early vocabularies is by size at various time points. The general picture of productive vocabulary has first words emerging at about 12 months of age, and by 24 months consisting of several hundred words. Dromi (1987) gathers some of the findings that were current at the time although she does not distill the data into a single summary figure, citing the differences in how researchers characterized a meaningful word and different sampling procedures. However, including her daughter Keren's data (the subject of Dromi (1987), learning Hebrew as her first language), and data from both Bloom (1973) and Braunwald (1978), the first productive word uses for all subjects occurred at 9-10 months of age. At 15(12)[1], Braunwald's subject (her daughter) had 50 words in her productive vocabulary, and in Dromi's case the subject had 102 words. By 17(23), the end of Dromi's study, the subject had 337 words and by 20(0) Braunwald's subject had 391 words. A sketch of these figures, together with the results of a study across multiple children by Fenson et al. (1994) (reported below), is shown in figure 4-1. It is immediately clear that one trait of vocabulary growth is its acceleration, even considering only three data points from each study.

The rapid increase in vocabulary size within a short period of time – the rate increase seen in the figure – is such a striking and commonly observed phenomenon that it often is given its own name: the word spurt, vocabulary explosion, or other similar name. Some have taken this feature as an indicator of a specific developmental change, such as an insight about how language functions, better categorization abilities, or otherwise. However, Bloom

[1] Shorthand for 15 months, 12 days of age.

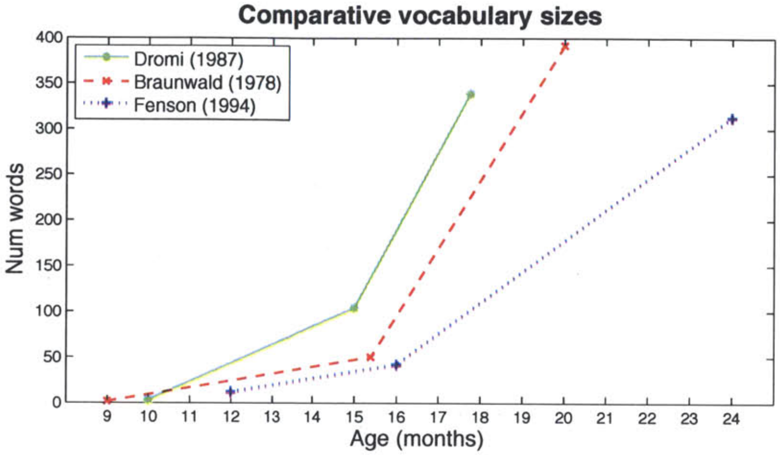


Figure 4-1: Vocabulary sizes from Braunwald (1978) and Dromi (1987), which are each obtained from one child. The data from Fenson et al. (1994) is the median vocabulary size across multiple children. Even these very few data points show a clear acceleration in vocabulary growth.

(2000) takes issue with this and the notion that there *is* a vocabulary spurt. For Bloom, one problem is that the criteria for a spurt have not been well defined. A sensible definition of a spurt is a sudden change from a slow to a fast rate of learning, rather than a gradual increase in learning rate. Ganger and Brent (2004) formalize these two models and test on data from 38 children, finding that the majority of children do *not* undergo a spurt, at least according to their definition. However, we do not find children's increase in learning rate any less impressive in spite of this finding. Children's productive vocabularies grow slowly at first, but accelerate in their second year of life. Indeed, acceleration must occur since, as Bloom calculates, to learn the conservative estimate of 60,000 words in an American or British high-school graduate's vocabulary requires learning approximately 10 words per day starting at 12 months of age. The word learning rates discussed thus far for the second year of life fall far short of this figure.

The work by Fenson et al. (1994) marked a major attempt at characterizing early language development across a very large sample of children. This was based on parental reports of more than 1,800 children from 8-30 months of age. Data from parents was collected via checklists developed for sampling infant and toddler language, referred to as the MacArthur-Bates Communicative Development Inventories (CDIs). At the time of the study, there was already a roughly 20 year history to the CDIs, so the instruments devised for the study were based on content and methodology with a solid foundation. This was a landmark study and provided a wealth of information on early lexical development. The authors found that children's receptive vocabularies grow well in advance of their productive vocabularies, with 8 month old children having a median receptive vocabulary of about 15 words, while the median productive vocabulary was less than 10 words at 12 months of age. By 16 months, the median receptive vocabulary was 169 words. However, Fenson et al. (1994) note the tremendous variability in children's receptive vocabulary sizes, particularly in later months, and they use the 10th and 90th percentiles as anchor points in reporting their findings. By 10 months, the 10th and 90th percentiles had receptive vocabularies of 11 and 154 words, respectively. By 16 months these figures had diverged even more, to 92 and 321 words for the 10th and 90th percentiles. Such variability partly drives the authors' critique of the notion of the "modal child" – the "mythical creature" that "passes through a predetermined sequence of stages, at a standard rate that is heavily constrained by maturational factors." This increase in variability holds for productive vocabulary as well. At 12, 16, and 24 months the 10th percentile showed no speech, fewer than 10 words, and under 100 words for each time point, respectively. For the same time points, the median had about 10, 40 and 311 words, and the 90th percentile about 26, 180, and 560 words. The authors note that, particularly toward the upper age ranges of the study (which ended at 30 months), the variability between the median and the 90th percentile started to decrease. This suggests that the 680 word checklist likely caused a "ceiling effect" in assessing vocabulary size for the children in the 70th percentile and above.

It is important to note that the values reported above do not capture the vocabulary growth for an individual child. It is not necessarily the case that a child in the 90th percentile at 12 months of age is also in the 90th percentile at other time points.

4.1.2 Vocabulary composition

Vocabulary size and growth rate are one way to characterize children's early lexical learning. But is there any pattern to the composition of children's early vocabulary? In fact, children's first words tend to be heavily weighted toward nouns and object names in general. Fenson et al. (1994) assessed the makeup of children's early vocabularies by considering those words that had been learned by 50% of children in a particular month, finding that by 16 months 52% of children's receptive vocabulary items were nouns while 18.8% were verbs. Of the first 188 words in children's *productive* vocabularies to reach the 50% criterion (which had all been learned by 22 months), 63.5% were nouns while only 8.5% were verbs, but by 30 months the percentage of nouns and verbs had shifted to 43.5% and 23.2% respectively. In fact, verbs were totally absent from the first 50 words in children's productive vocabularies.

Gentner and Boroditsky (2001) consider children's early vocabularies across different languages, finding that nouns are favored even in "verb-friendly" languages such as Korean, Mandarin and Italian. They suggest that names for animate beings seem to provide entry points to language, but also find that the accessibility of verbs in the input influences when verbs enter into children's vocabularies. For example, languages with many verb-final constructions are considered more "verb-friendly". This observation, that nouns are favored even in verb-friendly languages, suggests underlying cognitive factors that may be shaping *some* aspects of lexical acquisition. They make the case that some concepts arise naturally from pre-existing cognitive and perceptual factors and are simply "named by language", while other concepts are shaped by language itself. With this "division of dominance" idea in mind, the authors argue that nouns are easier to learn since the conceptual substrate is more readily accessible, but as the child enters into language linguistic factors can begin shaping conceptual development, and that there is a particular division between cognitive and linguistic dominance in how concepts develop. Further discussion on the contrast between nouns and verbs and related conceptual factors can be found in (Gentner, 2006).

4.1.3 Communicative development

Word learning is a major part children's early language acquisition. But although the focus of this work is on lexical development, there is much more to learning language than learning the words. To be competent language users, children must begin to productively combine words into more complex utterances. In (Dromi, 1987), the study spanned a total of 32 weeks, ending one week after the author noted that "Most of her new expressions are two- and multi-word utterances. It happened quite suddenly." (Dromi, 1987, p. 69). Dromi's daughter Keren was 17(16) when she noted the sudden prominence of multi-word utterances. In a 30 minute audio recording consisting of 279 utterances, 36% were multi-word expressions. Dromi makes the case for a specific one-word stage in early language development, although Snow (1988) argues that the conception of singular and major transition points may be wrong. Instead, she suggests thinking of cognitive development as a rope woven together of multiple strands, each strand progressing on its developmental trajectory but not all synchronously. Regardless, the onset of combinatorial speech is a major feature of children's second year of life. Fenson et al. (1994) find that by 22 months of age, more than 50% of parents report that their children are "often" combining words. Using a separate measure of sentence complexity as a proxy for children's mean length of utterance (MLU), children's median MLU reaches approximately 1.5 at 21 months of age, further evidence supporting combinatorial word use as children approach their second birthday.

4.2 Word learning in the Speechome Corpus

The developmental record in the Speechome Corpus spans the child's 9–24 month age range, capturing first words up to multiword speech. With more transcribed speech and a higher density of recordings than any prior study of a single child, what does early vocabulary growth look like?

4.2.1 Assessing vocabulary

Dromi (1987, p. 67) points out that receptive vocabulary is difficult to assess in any case study, and the same is true for the Human Speechome Project. Aside from administering word comprehension tests (effectively outside of the naturalistic, observational framework), assessing receptive vocabulary would likely depend on observing caregiver word uses in the child's presence and interpreting the child's behavior, requiring manual review and careful interpretation. However, an interesting semi-automatic approach to studying comprehension of a small set of selected words was undertaken in (Tsourides, 2010). This work focused on caregiver uses of words in the Speechome Corpus prior to the child's productive use of those words. Tsourides (2010) analyzed video of the child when the target word was uttered by caregivers, looking for characteristic motion patterns indicative of word comprehension. Essentially, the goal was to look for the naturalistic analogue of the "head-turn" experiments usually performed in controlled experimental settings. However, an extensive analysis of word comprehension in dense, longitudinal data remains a significant challenge.

The situation is more promising for studying productive vocabulary. While our study shares some similarities with prior diary studies, it obviously differs in many important respects. In (Dromi, 1987), the child's language use was effectively annotated directly by the experimenter, although some audio and video data was captured and later annotated. Our sampling process of near continual recording is more theoretically neutral, separating data collection from interpretation and analysis, but both the cost and the benefit is the large amount of transcribed speech to sift through in identifying the child's productive vocabulary.

One methodological issue that must be resolved is what constitutes a productive word use. For (Dromi, 1987), a new comprehensible word was admitted into the lexicon if its consistent sound sequence was repeated 3 times in various contexts, using the date of the third occurrence as the acquisition date. Note that a comprehensible word is not necessarily a conventional word, but rather a sound sequence that both speaker and hearer related to the same real-world referent. The context of the child's speech, and a familiarity with the texture of child-caregiver communication patterns are clearly important factors in determining what constitutes a word. However, an important point is that analysis of a word's meaning was *not* a requirement in assigning a word to the lexicon.

4.2.2 Word births

Although we were not originally familiar with the details of Dromi's work, we adopted very similar criteria for determining a first productive word use. Our intuition was that comprehensible, communicative speech, with specific reference and consistent phonology, should count as a word. Annotators transcribed the speech with this in mind, and with the transcripts as our starting point we sought to determine the timeline of the child's productive vocabulary. When we detected a new word in the child's productive vocabulary we called it a "word birth", effectively marking the age of acquisition (AoA) for each word. Although "word birth" is not a conventional term, we coined it to make clear that the detection of a new word in transcripts is an estimate of the true age of acquisition, but also for the more poetic connotation of a dynamic and growing language faculty.

4.3 Detecting word births

One way to identify the child's early vocabulary is as the union of all word types in child-produced utterances. By associating the timestamp of the utterance containing the first occurrence of each word, a word birth timeline can be constructed. Unfortunately, this is a very poor estimate of both vocabulary and AoA, since it fails to take into account two practical issues: annotation errors and corpus sampling characteristics.

Annotation errors are unavoidable for any large-scale annotation process. With millions of utterances transcribed by dozens of annotators, mistranscribed segments are inevitable. More significantly, since most speaker annotations are produced automatically, errors in speaker annotations must also be taken into account. However, with the above method an adult utterance mistakenly attributed to the child could potentially add many spurious words to the child's estimated vocabulary.

The second issue, that of corpus sampling characteristics, relates to how faithfully the corpus represents the subject of investigation. If significantly more data is recorded or annotated in particular weeks or months, then more word births may be identified at those times.

Words that are learned earlier may not be observed if their first productions are in sparsely sampled time ranges. The overall recording density varied as a function of the family's vacation plans and need for privacy, but in general the recording density is significantly beyond prior studies (further discussed in section 3.3). Furthermore, since analysis depends on annotated data, annotation density is a more important consideration. In `corpus08` the annotations were non-uniform and subsampling was necessary to balance the density before identifying words births. Since then, days have been selectively transcribed and `corpus12` is much more balanced. However, annotation errors may still pose problems for some analyses. In particular, while many analyses are robust to a small number of speaker ID or transcription errors, the above method for identifying word births is not.

4.3.1 First occurrence above threshold

Our first analysis of the child's productive vocabulary, described in (Roy et al., 2009), was based on `corpus08`. For technical reasons, this early version of the corpus was non-uniformly annotated as figure 3-14(a) makes apparent. Word births were identified by first subsampling to balance the data, and selecting only those utterances attributed to the child with very high speaker ID confidence. Words were normalized by hand to combine morphological variants (eg. plurals and gerunds). Initially, we required multiple uses of a word in utterances above threshold, but this eliminated too many low frequency words. Instead, candidate word births were reviewed manually. Figure 4-2 shows the number of word births over time in this data. The "shark fin" shape of word births by month was surprising, but stable across different random corpus subsamples and under different word birth identification strategies.

The first occurrence above threshold method was a good starting point for `corpus08`, but there is clearly room for improvement. One problem with the method was that, by filtering out utterances below a very high speaker ID threshold, words that are rarely but productively used by the child may not be detected. Furthermore, words that are detected may not be identified at their earliest occurrence and can be shifted later. Finally, with significantly

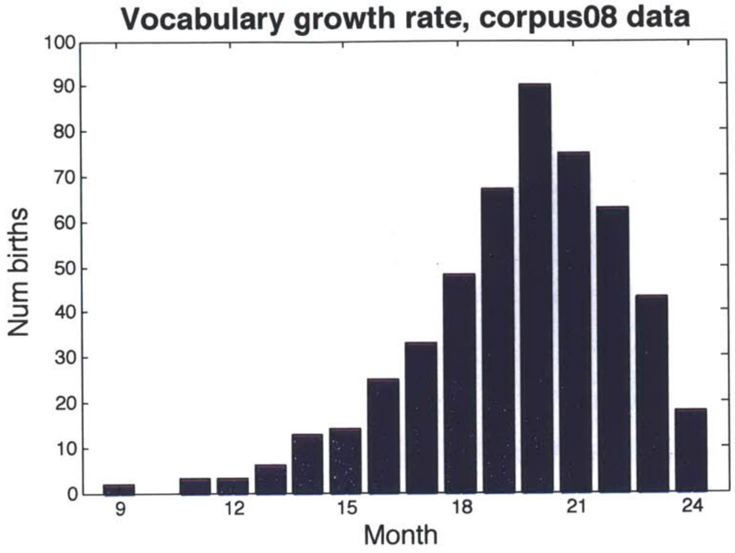


Figure 4-2: Word births by month based on `corpus08`, with a distinctive "shark's fin" shape.

more data in `corpus12` and correspondingly more annotation errors, more robust methods for word birth identification were required.

4.4 Robust word birth estimation

The two kinds of annotation errors described above can lead to errors in assessing the child's early vocabulary. One way to view a speaker ID error is as choosing the wrong label from a small set of candidates, while a transcription error is the result of choosing the wrong word (or words) from a much larger set of possible choices. In this sense, speaker ID errors are easier to correct and with a fully automatic speaker identification system the error rate can be quantified. On the other hand, transcriptions are human generated and at some level must be taken as ground truth. For these reasons, we focus on modeling speaker ID errors in devising a better method for identifying word births.

To begin, we consider all utterances attributed to the child at or above a particular speaker ID confidence threshold t. A threshold of $t = .4$ preserves 90% of the child's true utterances (the yield rate) at a false positive rate (FPr) of about .05. In this analysis, a false positive is a non-child produced utterance attributed to the child. Yield rate and false positive rate are described in section 3.1.4. The candidate child utterances are then tokenized and normalized via a manually generated mapping, reducing alternate spellings, plurals, gerunds and some common misspellings to a canonical root form, resulting in 6064 words. Next, words that are uttered 2 or fewer times by the child and 5 or fewer times overall are removed, leaving a set of 2197 candidate words. This is well beyond the expected vocabulary of a two-year old child, and at the limits of what could be manually reviewed, but a reasonable starting point for statistical methods.

One subtle point worth mentioning is that just as adult words may be misattributed to the child, child words may also be mistaken as adult speech or fall below the confidence threshold, excluding them from analysis. Such "false negatives" could artificially reduce estimates of the child's vocabulary size, just as false positives artificially inflate it. When estimating overall vocabulary size, it may be reasonable to assume the child knows some words that are never observed. In Mayor and Plunkett (2011), the authors seek to account for these missing words when estimating infant and toddler vocabulary using the MacArthur-Bates CDI scores. They find that the scores underestimate true vocabulary to a greater extent for older children and at higher CDI scores, and offer techniques for correcting vocabulary size estimates. However, in this work we are interested in the words themselves, and we are only willing to include a word in the child's lexicon if there is positive evidence for it.

Figure 4-3 shows the number of occurrences of the word "star" grouped by month. Gray bars represent the overall occurrence counts of the word, and red bars those attributed to the child at a speaker classification confidence greater than .4. From this plot, can we infer whether – and when – the child first produced the word? The "first occurrence" strategy described above would suggest the child's first use was at 9 months of age (May, 2006) since there are 3 occurrences in that month. However, this represents only about 4% of the 71 overall occurrences in month 9. Does this constitute sufficient evidence for a word birth? The following sections present several statistical approaches to identifying word births.

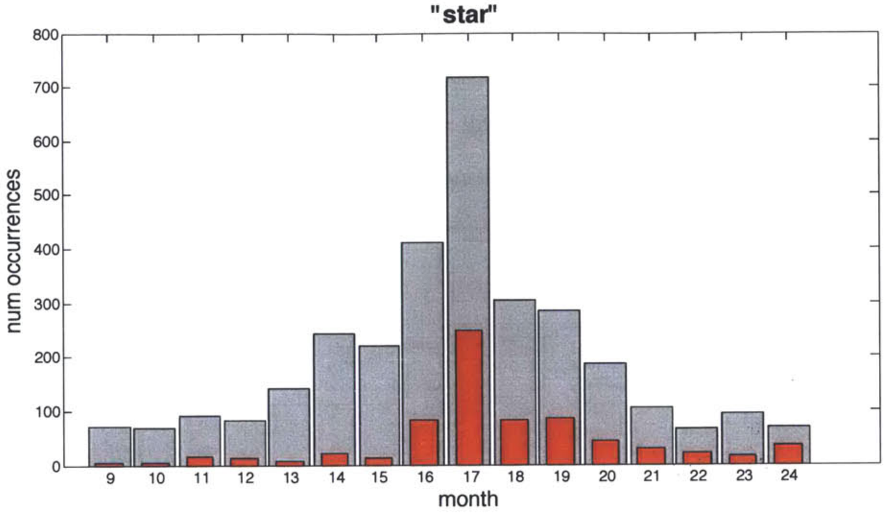
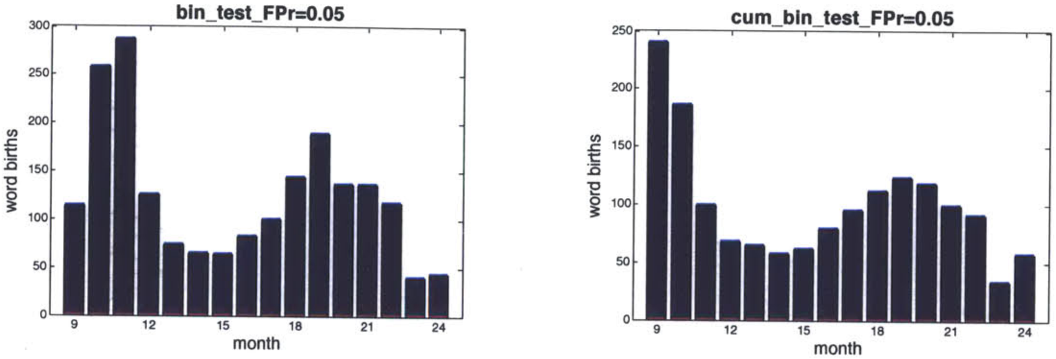


Figure 4-3: Usage counts of the word "star" by month. Gray bars indicate the overall number of uses of the word independent of speaker, while red bars are those uses attributed to the child at a speaker ID confidence threshold of .4.

4.4.1 Binomial models

Given the error rate of the speaker classification system, as described in section 3.1.4, what is the chance that all child uses of the word "star" are false positives? Under the null hypothesis that the child did *not* know the word "star", and that N occurrences of the word are drawn independently and from an identical underlying distribution, the number attributed to the child is a binomial random variable $X \sim \text{Binom}(N, \text{FPr})$. This yields an average of $\text{FPr} \cdot N$ false positives, and more importantly, a significance test for whether k child uses of the word are above chance. The first month in which the null hypothesis can be rejected at significance $\alpha = .05$ is month 11. Interestingly, it is only in month 16 (December, 2006) that the null hypothesis is rejected in all subsequent months. Under manual annotation, the word birth for "star" occurred in month 16.

We denote this test as `bin_test_FPr=.05`. Applying it to all 2197 words yields 1974 candidates at $\alpha = .05$ and 1718 at $\alpha = .01$. For each of the 1974 words in which the null hypothesis is rejected in some month, figure 4-4(a) shows a histogram of the first month in which the word is above chance. This test overestimates the child's vocabulary, and also



(a) Binomial test of child word counts for each month.

(b) Binomial test of cumulative child word counts for each month up to 24 months.

Figure 4-4: Histogram of word births by month testing the null hypothesis that child labeled word uses are the result of speaker ID false positives.

identifies far too many word births in early months. One issue is simply that each month is treated separately, and the test does not account for the fact that once a word is learned, it is very likely to remain in the child's lexicon in subsequent months.

A slight variation on this test, denoted as `cum_bin_test_FPr=.05`, is shown in figure 4-4(b). Here, a word's count at each month is the cumulative number of occurrences from that month up to 24 months. Applying this test and identifying the first month where the null hypothesis is rejected at $\alpha = .05$ results in 1579 word births, but has the same problem as the basic binomial test. The following table summarizes the number of words identified.

Method	Num words identified
`bin_test_FPr=.05`	1974
`cum_bin_test_FPr=.05`	1579

4.4.2 Testing goodness-of-fit and homogeneity

One issue with the above hypothesis testing scheme is that word births are checked individually by month. Whether or not a word is considered "learned" can vary by month, irrespective of the natural continuity from one month to another.

Test `G_test_FPr=.05` takes another perspective, which is that if a word is never learned all occurrences attributed to the child will be false positives, and these counts should track the overall number of occurrences by month assuming the speaker ID false positive rate remains constant. This can be tested with a χ^2 test or a more accurate likelihood ratio test (Rice, 2007, p. 341), sometimes called the G-test. The G-test statistic is defined as $G = 2\sum_i o_i \log \frac{o_i}{e_i}$ where o_i are the observed counts and e_i are the expected counts, and follows a χ^2 distribution with the appropriate number of degrees of freedom. In this case, for a particular word such as "star" the observed counts o_i for each month are as follows:

month	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
child	3	3	15	13	7	21	13	84	249	82	86	44	29	22	16	34
adult	68	65	75	70	135	222	206	328	467	222	200	142	77	44	77	36

Summing down each column yields $N_{\cdot i}$, the total occurrences for month i, and at FPr $= .05$ the expected child counts $e_{1i} = \mathrm{FPr} \cdot N_{\cdot i}$ and the expected adult counts $e_{2i} = (1 - \mathrm{FPr}) \cdot N_{\cdot i}$. The only issue with both the χ^2 and G-test is the usual requirement that $e_i \geq 5$. For many words, this condition is not satisfied and cells must be combined. For particularly rare words, the χ^2 and G-tests may not be appropriate, and an exact test should be used. Nevertheless, applying the G-test at $\alpha = .05$ leads to rejecting the null hypothesis for 748 words (ie. 748 word births detected)[2]. The test cannot be applied for 1357 words and fails to reject the null hypothesis for 92 words.

An alternative to testing the null hypothesis of a specific false positive rate is just to test whether the child and adult occurrences are drawn from the same underlying multinomial distribution. Assuming the child did not learn the word, then these two distributions should be homogeneous. We denote this test `G_test_homogeneous`. The expected count for speaker j at month i is $\epsilon_{ji} = \frac{N_{\cdot i}}{N} \cdot N_{j\cdot}$ where $\frac{N_{\cdot i}}{N}$ is the overall proportion of word occurrences for month i and $N_{j\cdot}$ is the total number of occurrences attributed to speaker j. The main difference here is that an additional parameter is estimated from the data, in place of using a predetermined false positive rate, and the degrees of freedom for the χ^2 distribution must

[2] With 2 rows and c columns, the number of degrees of freedom is df $= 2c - 1 - p$ where the number of estimated parameters $p = c - 1$, so df $= c$. There is one more degree of freedom than a typical contingency table since we are not estimating FPr (which, under the null hypothesis, would be the marginal probability for the child's row.)

be reduced appropriately. The result of this test at level $\alpha = .05$ is that the null hypothesis of homogeneity is rejected for 817 words (817 word births detected). The test cannot be applied for 1249 words and fails to reject the null hypothesis for 131 words.

The table below summarizes the number of word births detected. For the words where the null hypothesis is rejected, it is possible in principle to find the month (or months) that most contribute to rejecting the null hypothesis. However, we do not pursue this analysis here, in part because some months are combined for rare words.

Method	Num words identified
G_test_FPr=.05	748
G_test_homogeneous	817

4.4.3 Generalized likelihood ratio test

A more sophisticated approach to detecting word births is to compare a null hypothesis against an alternative hypothesis. As before, we consider the null hypothesis that the child did not learn the word and that child-attributed occurrences are false positives. The number of such false positives in month j is a random variable $X_j \sim \text{Binom}(N_j, p_0)$, where we assume that p_0 does not vary by month. The alternative hypothesis of a word birth at month i divides the data into two regimes: before and after the word birth. Prior to the word birth, for months $j < i$, $X_j \sim \text{Binom}(N_j, p_1)$. After the word birth, for months $j \geq i$, we have $X_j \sim \text{Binom}(N_j, p_2)$. We assume that p_1 and p_2 characterize their respective regimes and do not vary by month.

With this framework, the likelihood of $X_1 \ldots X_{16}$ under the null hypothesis is

$$L_0 = \prod_{j=1}^{16} \Pr(X_j; N_j, p_0)$$

and under the alternative hypothesis of a word birth at month i,

$$L_i = \prod_{j=1}^{i-1} \Pr(X_j; N_j, p_1) \prod_{j=i}^{16} \Pr(X_j; N_j, p_2)$$

The maximum likelihood parameter estimate for $p_0 = \sum_j X_j / \sum_j N_j$. The maximum likelihood estimates for p_1 and p_2 are the corresponding averages for each regime assuming a word birth at month i. However, we require that $p_2 \geq p_1$ since the alternate hypothesis assumes that the false positive rate of the speaker ID does not change, and that child-attributed occurrences include both true and false positives.

Among the set of alternate hypotheses indexed by i, the most likely month for a word birth is $i^* = \arg\max_i L_i$. This is also the maximum point for the log likelihood ratio function $l_i = \log L_i - \log L_0$ as L_0 is a fixed quantity for each word. Since the null hypothesis is nested within the class of alternate hypotheses, we can employ the likelihood ratio test to determine whether the null hypothesis should be rejected in favor of a word birth at month i^*. If the null hypothesis is true, then the test statistic $W = 2l_{i^*} = -2(\log L_0 - \log L_{i^*})$ has a χ^2 distribution with 1 degree of freedom, owing to the fact that the alternate hypothesis depends on one more parameter than the null hypothesis (p_1 and p_2, rather than just a single parameter p_0.)

Applying this test, denoted as `bin_split_LLR`, to all candidate words at $\alpha = .05$ yields 1375 word births (repeated in the table below.) For the word "star", the word birth is detected at month 16 (December, 2006), which corresponds to the manually identified birth date for this word. The number of word births by month is shown in figure 4-5. Although 1375 words births is significantly more than expected, the word birth histogram is more reasonable, with very few word births in early months and many more in later months. However, the decrease after month 19 is puzzling and will be discussed in the following sections.

Method	Num words identified
`bin_split_LLR`	1375

4.4.4 Word Birth Browser

The prospect of full manual word birth annotation is daunting. Without the statistical methods presented above, manual annotation could require listening to many thousands of

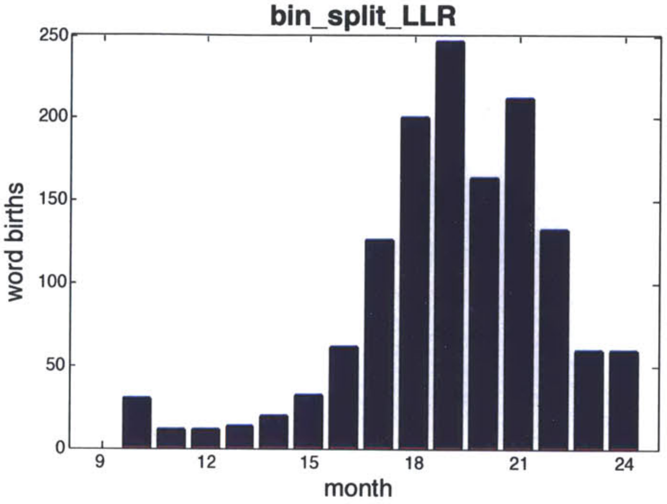
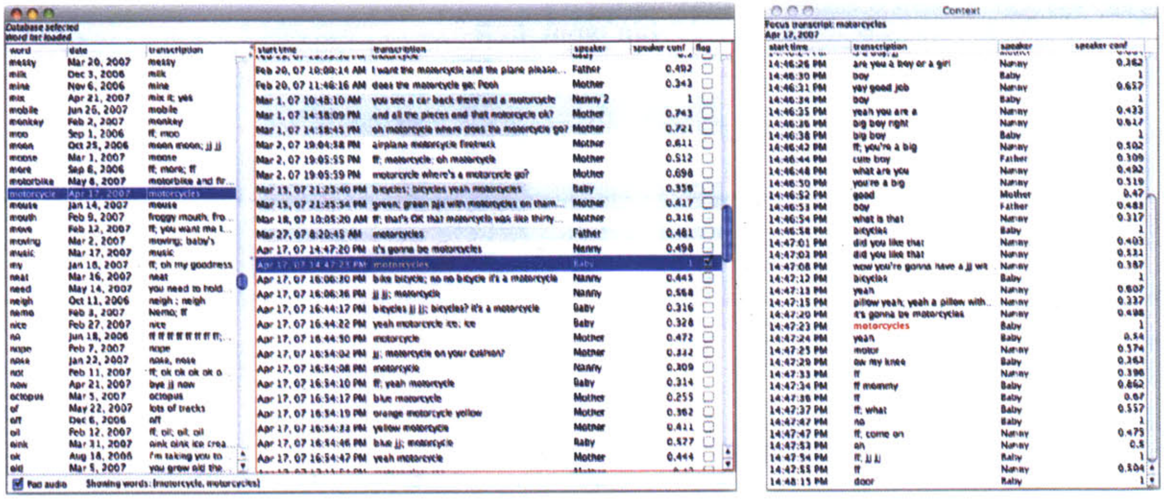


Figure 4-5: Histogram of word births by month, detected using the generalized likelihood ratio test.

utterances to find the first time the child used a particular word. For the very frequent words where there is little doubt that the child learned the word, many utterances containing the word would need to be reviewed to find the word birth date. For rare words there are fewer utterances to consider for each word, but there are many rare words. This is a reflection of the Zipfian word frequency distribution common to many text corpora.

The above statistical methods provide a useful starting point in identifying the child's early vocabulary, but since much of the analysis in this thesis depends on an accurate picture of the child's vocabulary growth timeline we perform a final, manual review of word births. The Word Birth Browser tool was built for this purpose, shown in figure 4-6. The left-hand window is the main window, consisting of two panels. The word birth panel on the left contains a list of candidate word births, along with the transcript that contained the corresponding birth[3]. Selecting a word birth in the word birth panel will update the right-hand panel with *all* transcripts containing that word (and its alternate forms) marking the transcript that contains the birth in red and centering it. By default, the transcripts in the right panel are sorted by time, but the list can also be sorted by speaker, speaker confidence and annotation checkbox. Nested sorting by multiple columns is supported. Selecting a

[3] The candidate transcript is usually a guess from those transcripts containing the target word that fall within the time range identified by the statistical estimate. In addition to estimating word birth month, word birth week was also estimated using the generalized likelihood ratio test of section 4.4.3.



Database selected
Word list loaded

word	date	transcription
messy	Mar 20, 2007	messy
milk	Dec 3, 2006	milk
mine	Nov 6, 2006	mine
mix	Apr 21, 2007	mix it; yes
mobile	Jun 26, 2007	mobile
monkey	Feb 2, 2007	monkey
moo	Sep 1, 2006	ff; moo
moon	Oct 25, 2006	moon moon; jj jj
moose	Mar 1, 2007	moose
more	Sep 6, 2006	ff; more; ff
motorbike	May 8, 2007	motorbike and fir...
motorcycle	Apr 17, 2007	motorcycles
mouse	Jan 14, 2007	mouse
mouth	Feb 9, 2007	froggy mouth, fro...
move	Feb 12, 2007	ff; you want me t...
moving	Mar 2, 2007	moving; baby's
music	Mar 17, 2007	music
my	Jan 18, 2007	ff; oh my goodness
neat	Mar 16, 2007	neat
need	May 14, 2007	you need to hold...
neigh	Oct 11, 2006	neigh ; neigh
nemo	Feb 3, 2007	Nemo; ff
nice	Feb 27, 2007	nice
no	Jun 18, 2006	ff ff ff ff ff ff ff ff;...
nope	Feb 7, 2007	nope
nose	Jan 22, 2007	nose, nose
not	Feb 11, 2007	ff; ok ok ok ok o...
now	Apr 21, 2007	bye jj now
octopus	Mar 5, 2007	octopus
of	May 22, 2007	lots of tracks
off	Dec 6, 2006	off
oil	Feb 12, 2007	ff; oil; oil, oil
oink	Mar 31, 2007	oink oink ice crea...
ok	Aug 18, 2006	I'm taking you to...
old	Mar 5, 2007	you grow old the...

start time	transcription	speaker	speaker conf	flag
Feb 20, 07 10:00:14 AM	I want the motorcycle and the plane please...	Father	0.492	□
Feb 20, 07 11:46:16 AM	does the motorcycle go; Pooh	Mother	0.343	□
Mar 1, 07 10:48:10 AM	you see a car back there and a motorcycle	Nanny 2	1	□
Mar 1, 07 14:58:09 PM	and all the pieces and that motorcycle ok?	Mother	0.743	□
Mar 1, 07 14:58:45 PM	oh motorcycle where does the motorcycle go?	Mother	0.721	□
Mar 2, 07 19:04:58 PM	airplane motorcycle firetruck	Mother	0.611	□
Mar 2, 07 19:05:55 PM	ff; motorcycle; oh motorcycle	Mother	0.512	□
Mar 2, 07 19:05:59 PM	motorcycle where's a motorcycle go?	Mother	0.698	□
Mar 15, 07 21:25:40 PM	bicycles; bicycles yeah motorcycles	Baby	0.358	□
Mar 15, 07 21:25:54 PM	green; green pjs with motorcycles on them...	Mother	0.417	□
Mar 18, 07 10:05:20 AM	ff; that's OK that motorcycle was like thirty...	Mother	0.316	□
Mar 27, 07 8:20:45 AM	motorcycles	Father	0.481	□
Apr 17, 07 14:47:20 PM	it's gonna be motorcycles	Nanny	0.498	□
Apr 17, 07 14:47:23 PM	motorcycles	Baby	1	☑
Apr 17, 07 16:06:30 PM	bike bicycle; no no bicycle it's a motorcycle	Nanny	0.445	□
Apr 17, 07 16:06:36 PM	jj jj; motorcycle	Nanny	0.568	□
Apr 17, 07 16:44:17 PM	bicycles jj jj; bicycles? it's a motorcycle	Baby	0.316	□
Apr 17, 07 16:44:22 PM	yeah motorcycle ice; ice	Baby	0.328	□
Apr 17, 07 16:44:50 PM	motorcycle	Mother	0.472	□
Apr 17, 07 16:54:02 PM	jj; motorcycle on your cushion?	Mother	0.332	□
Apr 17, 07 16:54:08 PM	motorcycle	Nanny	0.309	□
Apr 17, 07 16:54:10 PM	ff; yeah motorcycle	Baby	0.314	□
Apr 17, 07 16:54:17 PM	blue motorcycle	Mother	0.255	□
Apr 17, 07 16:54:19 PM	orange motorcycle yellow	Mother	0.362	□
Apr 17, 07 16:54:33 PM	yellow motorcycle	Mother	0.411	□
Apr 17, 07 16:54:46 PM	blue jj; motorcycle	Baby	0.577	□
Apr 17, 07 16:54:47 PM	yeah motorcycle	Mother	0.444	□

☑ Play audio Showing words: [motorcycle, motorcycles]

Context

Focus transcript: motorcycles
Apr 17, 2007

start time	transcription	speaker	speaker conf
14:46:26 PM	are you a boy or a girl	Nanny	0.362
14:46:30 PM	boy	Baby	1
14:46:31 PM	yay good job	Nanny	0.657
14:46:34 PM	boy	Baby	1
14:46:35 PM	yeah you are a	Nanny	0.433
14:46:36 PM	big boy right	Nanny	0.617
14:46:38 PM	big boy	Baby	1
14:46:42 PM	ff; you're a big	Nanny	0.502
14:46:44 PM	cute boy	Father	0.309
14:46:48 PM	what are you	Nanny	0.492
14:46:50 PM	you're a big	Nanny	0.519
14:46:52 PM	good	Mother	0.47
14:46:53 PM	boy	Father	0.483
14:46:54 PM	what is that	Nanny	0.317
14:46:58 PM	bicycles	Baby	1
14:47:01 PM	did you like that	Nanny	0.403
14:47:03 PM	did you like that	Nanny	0.531
14:47:08 PM	wow you're gonna have a jj wit...	Nanny	0.387
14:47:12 PM	bicycles	Baby	1
14:47:13 PM	yeah	Nanny	0.607
14:47:15 PM	pillow yeah; yeah a pillow with...	Nanny	0.337
14:47:20 PM	it's gonna be motorcycles	Nanny	0.498
14:47:23 PM	motorcycles	Baby	1
14:47:24 PM	yeah	Baby	0.54
14:47:25 PM	motor	Nanny	0.374
14:47:29 PM	ow my knee	Baby	0.363
14:47:33 PM	ff	Nanny	0.398
14:47:34 PM	ff mommy	Baby	0.862
14:47:36 PM	ff	Baby	0.67
14:47:37 PM	ff; what	Baby	0.557
14:47:47 PM	no	Baby	1
14:47:47 PM	ff; come on	Nanny	0.475
14:47:53 PM	oh	Nanny	0.5
14:47:54 PM	ff; jj jj	Baby	1
14:47:55 PM	ff	Nanny	0.504
14:48:15 PM	door	Baby	1

Figure 4-6: Word birth browser main window (left) and context window (right). The main window shows the list of vocabulary items and all transcripts containing the selected word. Selecting a transcript in the main window updates the context window to show all transcripts in a time window surrounding the selected transcript. The audio can be played by pressing the TAB key.

transcript in this list and pressing TAB will play the audio segment. The checkbox column provides a very basic annotation capability. Annotations can be saved and loaded from a file. Selecting a transcript also updates the context window, shown on the right. This window shows all transcripts that fall within a short time range before and after the selected transcript.

The Word Birth Browser was developed in Java, retrieving data from a local `Sqlite3` database containing a version of the corpus. The word births are loaded from a file. Alternate word forms may also be loaded, so that when a target word is selected in the word birth panel, transcripts containing that word or alternate forms are retrieved.

The Word Birth Browser was written in a short period of time, and rather than implementing extensive workflow and annotation capabilities a process was developed to manage annotation. A shared Google Document contained the list of all candidate word births, and the words were divided into groups and assigned to the transcribers who took part in this project. Each transcriber would listen to the candidate in their list using the Word Birth Browser, and if the segment contained the child using the word they would search for

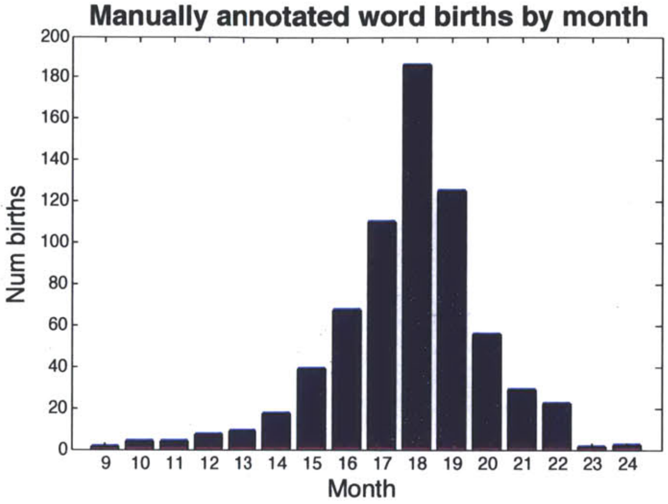


Figure 4-7: Word births by month for manually checked and annotated word births.

earlier examples. In some cases, the word birth could not be confirmed either before or after the candidate word birth and the word would be rejected entirely. The final word birth annotation was made by selecting the appropriate segment checkbox. This process went fairly quickly. The first round of annotations took only a few weeks with four annotators, including time spent in training and discussions as we developed the process. In a second round some months later, four annotators revised some of the word births over the course of about two weeks. Through the manual annotation process, 669 word births were identified.

Method	Num words identified
`manual-ann`	669

Word births by month are shown in figure 4-7. The general picture of vocabulary growth remains the same: words are added to the vocabulary at an increasing rate up to 18 months of age, followed by a decrease in growth rate.

4.4.5 Birth Estimation Evaluation

The word birth estimates can be evaluated with respect to the human annotated word births in two ways. First, all methods discussed above yield a list of vocabulary items that the child may have learned. Second, some of these methods provide an estimate of when the word was learned, which can be compared with the human annotated AoA if available.

To evaluate the word births detected (but not their AoAs) we calculate a similarity coefficient between word sets A and B as

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

This ranges between 0 (no words in common) and 1 (identical sets), and is often called the Jaccard similarity coefficient. Table 4.1 summarizes the results of this comparison. An important caveat, however, is that the statistical estimation methods evaluated the same set of 2197 initial words while the set of words evaluated by the annotators was a smaller set. The words the annotators worked with were derived from an accumulation of word lists generated over the course of several years, as different researchers worked with the data (eg. Shaw (2011) developed different methods for word birth identification.) Therefore, the vocabulary set similarity measures should be taken with a grain of salt, since it is possible that the statistical methods detected true word births that were actually missed in the manual annotation process. Ideally, the manual annotation process would be applied to the same initial set of words for a more accurate comparison. Unfortunately, this is impractical given the time and costs of further word birth annotation. Nevertheless, the results are still useful in assessing the estimators, since a good estimator should detect true word births and result in a high overlap with the manually identified set.

The second evaluation of interest is for those estimators that estimate word birth date. For the detected word births that overlap with the manually annotated word births, we can calculate the difference between the estimated AoA for a word and the manually annotated AoA for that word. This difference is essentially the "error" of the estimate, although we refer to it as an "offset" since, in our experience, all annotations are estimates whether by human or machine.

For a given estimation method and the set of words w_i in common with the manually annotated list, the offset $\delta_i = t_i^{\text{estim}} - t_i^{\text{manual}}$ is calculated, where t_i^{estim} is the estimated birth

Estimation method	Num words identified	Num words in common	Set similarity	avg(δ_i)	stdev(δ_i)	avg($\|\delta_i\|$)
bin_test_FPr=.05	1974	647	0.323	-2.76	4.27	3.79
cum_bin_test_FPr=.05	1579	610	0.370	-4.88	3.79	5.16
G_test_FPr=.05	748	493	0.528	—	—	—
G_test_homogeneous	817	588	0.648	—	—	—
bin_split_LLR	1375	628	0.440	0.67	2.41	1.71
manual-ann	669	—	—	—	—	—

Table 4.1: Evaluation of the word birth estimation methods against the manually annotated word births. The number of words identified by the method, the number of words in common with the manually annotated word births, and the set similarity are summarized in the first columns. Note that each method started with the same initial set of 2197 words, except for the set of manually annotated words, which have accumulated over several years of work. The last three columns summarize the AoA estimate offsets relative to the manually annotated AoAs. This only applies for the three methods that identify word birth dates.

date and t_i^{manual} is the manually annotated birth date for word w_i. The offset units are in months. We calculate the mean, standard deviation and mean absolute value over $\{\delta_i\}$ to quantify the AoA estimates, which are also summarized in table 4.1. The first two methods (bin_test_FPr=.05 and cum_bin_test_FPr=.05) have negative means, indicating that the estimates are earlier than the manual annotations, with large standard deviations and mean absolute offsets. This should be expected simply by inspecting their AoA histograms – there are far too many word births in the earliest months. The best method is bin_split_LLR, with AoA estimates shifted slightly later, on average, and with an average absolute offset of 1.71 months.

More details on bin_split_LLR are shown in figure 4-8, which shows the estimated and manually annotated AoAs for the words in common overlaid on the originally estimated birth histogram. A scatter plot of the manually identified vs. estimated AoA for each word is shown in figure 4-9(a), and a histogram of the offsets in figure 4-9(b). The histogram illustrates that most of the offsets are near zero, but the right tail is heavier (and hence the positive mean). In addition to the standard deviation, we calculate the mean absolute offset to help summarize this histogram since it seems more peaked than a Normal distribution.

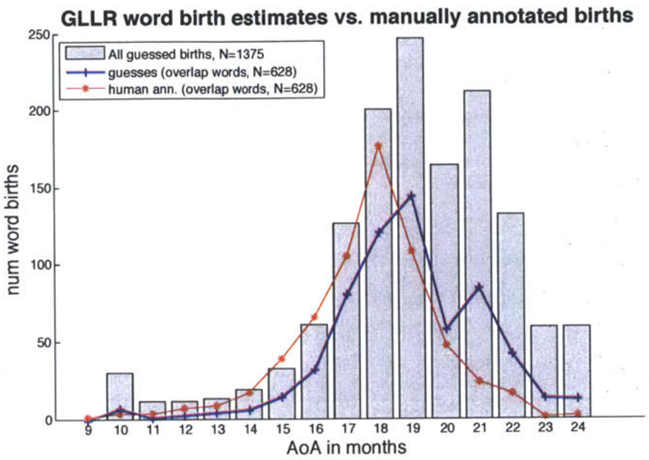
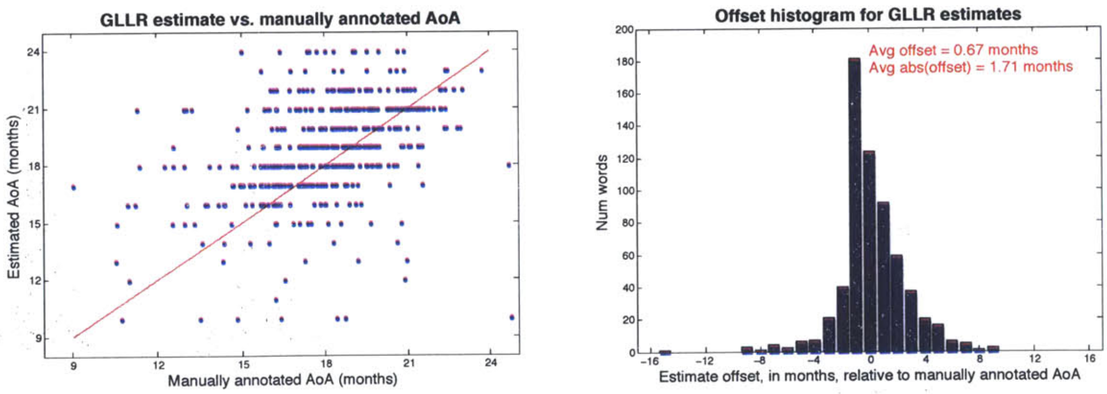


Figure 4-8: Gray bars show the histogram of word births by month detected using the generalized likelihood ratio test, which is the same as that shown in figure 4-5. The blue line with + marks shows the same word birth estimates, but for the subset of words that overlap with the manually annotated births. The red line with * marks shows the human annotated births for this same subset of words. In general, the bin_split_LLR method detects word births somewhat later since the blue line is shifted to the right of the red line.



(a) The bin_split_LLR estimates against the manually annotated AoAs for the set of common words.

(b) Histogram of the deviation of the bin_split_LLR estimates relative to the manually annotated AoAs.

Figure 4-9: Comparison of the bin_split_LLR estimated word birth dates to the manually annotated word birth dates, for the 628 words that are common to both sets.

4.5 Interpreting the child's vocabulary growth

When the child's word birth curve first came to light in (Roy et al., 2009) it was surprising, but after transcribing significantly more data and developing more reliable methods, the picture remains much the same. From our first study to the present one, the peak of vocabulary growth has shifted from 20 months to 18 months of age (compare figures 4-2 and 4-7), but the child's vocabulary growth still has an overall "shark's fin" shape. Word births occur at an accelerating rate followed by a sharp decline.

Up until the peak of the curve at 18 months, word births accelerate smoothly. Whether such a trajectory should be considered a "vocabulary spurt" largely depends on how it is defined, as mentioned above. However, this is not simply a matter of definitions but also of what accelerating vocabulary growth might indicate. Gopnik and Meltzoff (1987) suggest the onset of a naming insight, where children discover that all things can and should be categorized, giving rise to a "naming explosion" as a rapid growth of object words. Unfortunately, their definition of the naming explosion is based on whether the number of new object words crosses a threshold, which Bloom (2000) understandably takes issue with. Goldfield and Reznick (1990) also question whether these thresholds really mark a sharp change in the rate of word-learning or simply a more gradual pattern of acceleration. But these authors are comfortable referring to an accelerated period vocabulary growth as a longer-term "surge", and in any case are more interested in how vocabulary composition relates to the rate of growth. They suggest that children whose vocabulary growth rate accelerates may be more noun-centric, while children with flatter vocabulary growth curves may take a more balanced approach, learning a more even distribution of nouns, verbs and other word classes. Clark (1995) contrasts data from her own diary study showing steady lexical growth with the accelerating growth in (Dromi, 1987), and discusses related phenomena such as the children's use of word combinations and phonological development. In our work, the subject is very noun-centric and his vocabulary growth rate certainly shows a sustained "surge".

Although not emphasized in their work, Goldfield and Reznick (1990) wonder what happens after the spurt or surge – whether accelerated lexical growth continues or stabilizes. Our data show a clear decrease in lexical growth rate, although growth does continue. The striking drop in growth rate is less well documented, largely because it may not be so dramatic – or observable at all – when considering aggregate vocabulary growth curves across many children. Fenson et al. (1994) point this out for their own study, noting that averaging different vocabulary growth functions may obscure non-monotonic patterns. In fact, although the median and higher vocabulary size curves do flatten out at later months (corresponding to slower vocabulary growth), they suspect this is an artifact of the 680 word checklist failing to accommodate children with larger productive vocabularies. The challenge in detecting word births and the limitations of using any checklist to fully capture a child's vocabulary suggest that prior studies may have underestimated children's true productive vocabulary sizes. To be fair, the (Fenson et al., 1994) study purposely selected a common subset of words to form their checklist – too large a list would be cumbersome for parents to use. Furthermore, their goal was not necessarily to identify a child's complete vocabulary but rather to develop a normative measure of lexical development. Nevertheless, these are two important points in studying lexical development: checklists may lead to underestimates of vocabulary size and averaging over multiple children may obscure non-monotonic lexical growth patterns for individual children. The authors suggest that longitudinal studies are needed to address such individual variability.

One prior longitudinal study that *does* show a striking drop in vocabulary growth rate is that presented in (Dromi, 1987). Figure 4-1 showed only 3 data points from Dromi's data, but the full curve looks remarkably similar to our own (shown in figure 4-7). This provides strong supporting evidence for such a "shark's fin" growth rate curve in individual children's vocabulary, but leaves the question of its origin unanswered. Dromi (1987) argues that early language acquisition is stage-like, and a distinct "one-word" stage is winding down in preparation for the child's transition into syntax. Dromi notes that in the weeks of decreasing vocabulary growth rate, her daughter seemed to be exploring the words she had already learned, refining their use, and generally consolidating the lexicon. We find this a compelling idea, and since our first analysis in (Roy et al., 2009) we have wondered whether the drop in word birth rate could coincide with an increase in the child's use of syntax. We explore this possibility at the end of the chapter.

Another possibility, and one that may not at first sound like a serious proposal, is that the child has learned the words he needs and productive vocabulary growth slows down as a result. It is hard to imagine that with 669 words the child's communicative needs are satisfied. Then again, the child has responsive caregivers and the range of activities in a 9–24 month old's life are limited. The introduction of a new toy, activity or other experience (such as going to the zoo) could contribute new words in the child's lexicon, but at a certain point the child's vocabulary may be sufficient for the activities of everyday life. Built into this idea is a notion of a communicative *need* that word learning satisfies. The implicit perspective is that language growth is not simply a developmental process running its course but a process driven in response to both internal and external forces. We will not push this argument further at present, but note that Bruner (1983), in his naturalistic, longitudinal study of two young children's social interaction with their parents, found that caregivers continually push children toward more sophisticated communicative behavior. For example, when the mother knew the child could use a word, she would begin to *require* its use in appropriate situations. For now though, we leave this as a vague but intriguing idea for further consideration.

Returning to the left side of the shark's fin – the period of accelerating growth – McMurray (2007) asked whether this well-known growth spurt could be a mathematical consequence of two basic conditions of word learning: that words are acquired in parallel, and that some words can be learned quickly while most require more time. The authors constructed a very simple computer simulation in which word difficulty is normally distributed, and each time step pushes each word closer to its acquisition threshold. Since there are many factors that may combine and contribute to a word's difficulty, the authors justify the normal difficulty distribution by the Central Limit Theorem. Under a variety of conditions, in both their original and later work (Mitchell and McMurray, 2008), the simulations yield a vocabulary explosion. The relevance of this finding is that an accelerating growth curve can be explained without appealing to a change in the learning process, if the right conditions hold. Of course, the force of this argument depends on having a reasonable model of the conditions. Mayor and Plunkett (2010) suggest a different model, but theirs leads to a very different result that does *not* show accelerating vocabulary growth, which they use to argue for an endogenous change in the learner.

The right side of the shark's fin – the period of decelerating growth – has received little attention, partly because observing it requires a unique kind of data. Dromi (1987, p. 113) suggests as much, noting that systematic daily recordings are needed to construct a detailed growth curve, and that as the child's lexicon grows it becomes harder to document. Could the observed decrease in growth rate be an artifact of sampling, as suggested by Bloom (2000, p. 43)? If so, it would have important consequences for theories of lexical acquisition.

4.6 Models of vocabulary growth

If one were to assess a child's vocabulary as only the words he used during a brief visit, it would almost certainly fall short of his true vocabulary. In a longitudinal study, the chance of observing any particular word depends on multiple factors, including when the word is learned, its usage frequency, and the number of "samples" – or tokens uttered – that are available. Could a model in which the true vocabulary is growing at a constant or increasing rate give rise to a shark's fin shape?

Suppose that at time t, the child's true vocabulary contains V_t words, and that k_t words are uttered at random from a distribution f_t over these words. We assume that observations are taken uniformly and when a word is uttered at least c times it can be considered observed. Then let O_t be the size of the observed vocabulary at time t. These are the rough outlines for a simple class of models for experimentation.

We consider three functional forms for V_t – constant, linear and exponential functions of t. The growth of V can be a random process, but for now we use simpler deterministic functions for vocabulary size. However, we do incorporate randomness into the child's production model, where a word is selected from V according to f, such as a uniform or Zipfian distribution. As V grows, f must change appropriately (hence f depends on t, as denoted by f_t.) For example, with a Zipfian distribution, old words can preserve their rank, ranks may increase, or all ranks may be reshuffled. For now we assume that ranks are preserved. Finally, the number of tokens uttered by the child k_t can vary with time or may be fixed, but we start with a fixed k. For conceptual simplicity, we refer to an interval as a "day" and run the simulation for 448 days. We choose 448 days since it divides evenly into 16 intervals of 28 days each, which can be thought of as "months". This is fairly close to the true recorded Speechome Corpus of 444 days over 16 months, and using these comparable names and units helps in linking the simulated data to the observed data.

Representative outputs from running this simulation are shown in figure 4-10. Each "day", $k = 5000$ tokens are randomly sampled from the current vocabulary according to f_t, here with f_t as $\mathrm{Zipf}(\alpha, V_t)$, with no reshuffling of vocabulary items – in effect, the vocabulary grows only by adding lower frequency words. The daily word type counts are grouped into "months", and a word birth is detected if the word count for that month is greater than or equal to a threshold $c = 2$. This is effectively the "first occurrence above threshold" detection process described above, and suitable for this simulation since we are not concerned with other noise sources (such as speaker ID or transcription errors.)

As might be expected, the number of word births only decreases with time for a constant vocabulary size. The odds of choosing an unseen word are high early on when no word births have been detected, but as more of the vocabulary is observed, the odds of choosing a previously unseen word decreases. For a linearly increasing vocabulary, whether the number of word births increases, decreases, or remains roughly constant depends on the total probability mass being introduced to the unseen set by the new vocabulary items. This depends on the usage probability of the new words, and for a Zipfian distribution where the new words are out on the tail, this probability is ever decreasing. As the middle figure shows, when $\alpha = 2.1$ the word birth rate decreases over time. On the other hand, for an exponentially growing vocabulary, enough new words are added to the vocabulary to counteract the power-law falloff in their probability, at least for the smaller $\alpha = 1.5$. However, an exponentially growing vocabulary clearly cannot be a long term phenomenon, since the same exponential growth curve extended for 2 more "months" would nearly triple the child's vocabulary.

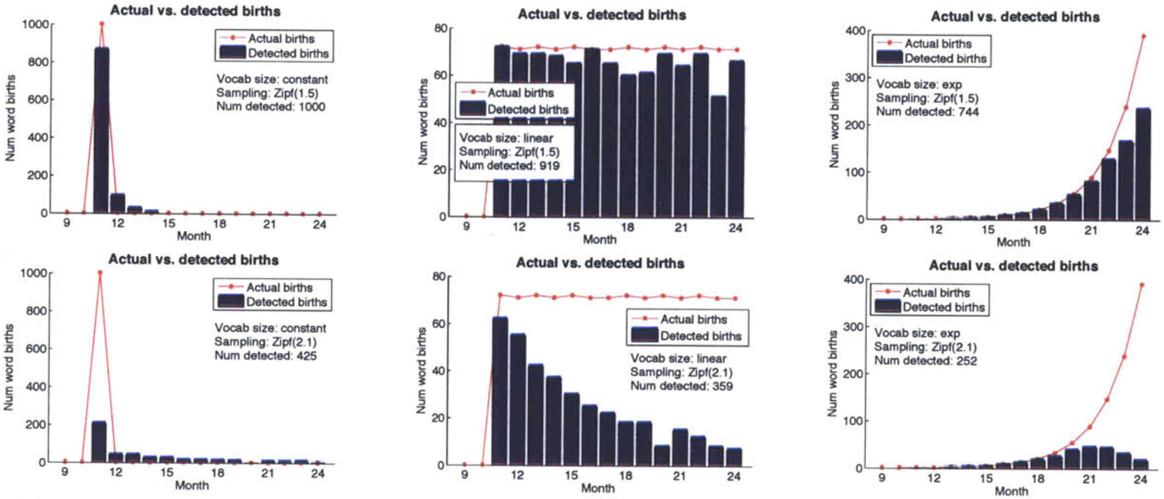
The main simulation of interest is the exponential vocabulary growth model with steep Zipfian falloff parameter $\alpha = 2.1$, shown in more detail in figure 4-11. This curve is interesting since, qualitatively, it shows an increasing and then decreasing word birth rate, the "shark's fin." However, here only 252 of the 1000 words in the child's "true" vocabulary are observed, a serious underestimate. Most simulations that yielded similarly shaped curves also severely underestimated the true vocabulary.

These simulations are useful in articulating and exploring how a growing vocabulary interacts with a stochastic usage model. The model could be extended first by considering a more sophisticated usage model, for example, one in which new words are not the lowest frequency words. Such an extension would be both more realistic but also promote detection of real word births. Another extension could be to consider a vocabulary that grows stochastically.

While it may be the case the disproportionately more word births are missed in later months, these initial explorations suggest that the shark's fin is probably not a sampling artifact. If the child's *true* vocabulary keeps growing at a fixed or increasing rate (ie. linear or exponential vocabulary in our models), it is unlikely that the number of *observed* word births will decrease except under very special, unrealistic conditions. Therefore, we argue that the word births we have detected in the Speechome Corpus, shown in figure 4-7, is an accurate reflection of the child's true vocabulary growth rate.

4.7 Combinatorial speech

If the drop in vocabulary growth rate is not a statistical artifact, as suggested in the previous section, what else could contribute to the "vocabulary implosion" observed? Before 19 months of age, the child has 444 words in his productive vocabulary. If word learning is partly fueled by "communicative need", does the decrease in vocabulary growth rate indicate that the child has achieved some level of communicative sufficiency at 18 months? Or does



(a) Constant vocabulary (starting at 11 months)

(b) Vocabulary with linear growth

(c) Vocabulary with exponential growth

Figure 4-10: Actual and detected word births for constant, linear and exponential vocabularies. Words are uttered at random according to a Zipfian distribution with parameter $\alpha = 1.5$ (top row) or $\alpha = 2.1$ (bottom row). A fixed $k = 5000$ tokens are uttered per day, and in all cases the true vocabulary is 1000 words by 24 months. The only model in which true vocabulary continues to grow while the observed vocabulary has a "shark fin" shape is the exponential + Zipf(2.1) combination (bottom right, detail in figure 4-11).

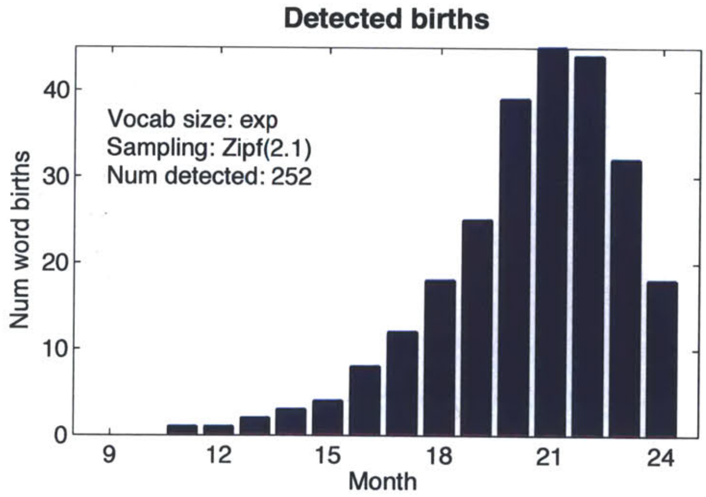


Figure 4-11: Detail on detected word births (actual word births not shown for better vertical scale) for the exponential growth model with Zipf(2.1) usage distribution. This produces an observed vocabulary with a shark's fin shape, but the combination of parameters to obtain this shape are extreme, significantly underestimating the child's true vocabulary at 252 out of 1000 words, and assuming a very steep falloff in word usage probability.

communicative growth transition from learning new words to combining words together in new ways?

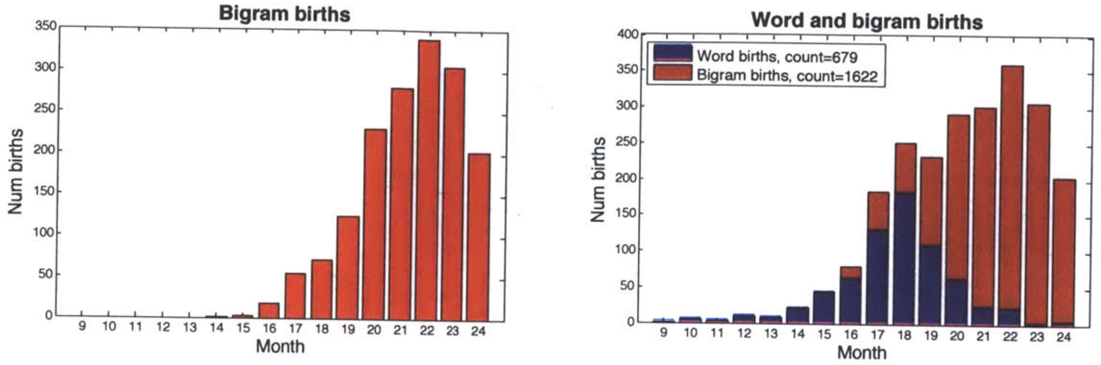
To explore this question, we used the same techniques for detecting word births to detect "construction births" – the first reliable occurrence of a particular word combination. We first identified all sequential pairs of words, or *bigrams*, uttered by the child at the same confidence level used for word births ($t = 0.4$), excluding multispeaker utterances. However, we made the following simplifying assumption: a particular bigram can only be considered if each of its constituent words has been acquired prior to the bigram date.[4]

In constructing the bigrams, `ll` and `ff` were removed from the utterance first since these are non-word types. Thus, the bigram "red car" would be extracted from the utterance "red ll car". However, any bigram that was a simple repetition (eg. "car car") was excluded, and finally, bigrams occurring two or fewer times were excluded.

This process yields a list of unique bigrams and a list of occurrence timestamps for each bigram. The timestamps for a bigram are then binned into monthly usage counts. The only difference between the count histogram for a bigram and the count histogram for a word is the starting month; for a bigram, the first possible month is the month containing the later of the two constituent word births. The `bin_split_LLR` test from section 4.4.3 is then applied to the histogram for each bigram to identify first whether there is sufficient evidence for a "bigram birth", and if so, the maximum likelihood birth month. A sample of the bigrams that the child learns are adjective-, verb- and preposition-noun combinations, such as "black car", "big truck", "see mama", "in chair". There are also many determiner-noun combinations, such as "the bed", "the telephone", noun- or pronoun-verb combinations such as "daddy read", "you throw", and so on.

Figure 4-12 shows the bigram birth curve by itself and combined with the word birth curve. There are several interesting points that can be noted here. The most striking point is that bigrams seem to "fill in" the gap left by the declining word birth rate curve, as the combined graph shows. The bigram birth rate also increases dramatically, with a decrease starting

[4] But see the discussion at the end of this section on whether words are really the constituent units that the child is learning.



(a) "Bigram" births

(b) Word births and bigram births

Figure 4-12: The number of detected "bigram births" by month, and bigram births stacked on top of word births by month. The decline in word birth rate is largely filled in by new bigrams.

in month 23. One possibility for this decrease is a detection problem in later months, but another possibility is that bigrams are being reused as part of longer constructions, such as 3-word long *trigram* sequences. One piece of evidence that suggests trigram births may fill in the gap after bigrams births decrease is found in our earlier study in which we calculated the child's mean utterance length (MLU) in words for each month (Roy et al., 2009). After month 22, the child's MLU increases past 2 and approaches nearly 3 words, an indicator of more and longer utterances, and possibly the source of new, unique three word combinations.

To push this analysis further, some care should be taken in determining the context in which a word or bigram is first used. For example, suppose the words "red" and "ball" are always and only used together as "red ball". Should this count as two word births or a single bigram birth? If later on the sequences "red car" and "green ball" are used, it suggests that "red" and "ball" are now functioning as constituent words. As Peters (1983) discusses, the fundamental units of language that children learn may not be words in their final adult form, and even if they are used they may not be analyzed into their constituent units until a later time. Clark (1995) also mentions some of her conditions for considering new word combinations.

Becoming a language user is about more than learning words; words can be combined together into more complex, expressive utterances. Vocabulary growth is part of overall linguistic and communicative development, and while vocabulary growth rate decreases other aspects of language development begin to accelerate.

4.8 Conclusion

Children begin producing their first words around their first birthday, and by two years of age often have productive vocabularies of several hundred words. Large-scale, naturalistic, longitudinal corpora that capture this time period present an opportunity to explore the details of lexical acquisition. We have undertaken such a study with the Speechome Corpus, but its complexity and scale present many challenges. Receptive vocabulary is particularly difficult to assess from transcript data, but the prospects are better for measuring productive vocabulary. By using appropriate noise-robust methods and tools on the Speechome data, we extract the child's productive vocabulary, which may help shed light on patterns of lexical acquisition.

In the Speechome Corpus, we find that vocabulary growth begins slowly and accelerates over time, consistent with earlier findings of both individual growth curves (Dromi, 1987) and aggregate data (Fenson et al., 1994). More surprisingly, we find that at a certain point the rate of lexical acquisition *decreases*, largely unobserved in aggregate data but consistent with the detailed, longitudinal study in (Dromi, 1987). We have taken care to consider the possible contribution of sampling artifacts, but feel confident in the final word birth timeline.

By 24 months of age, the child had learned 669 words. He learned these words through exposure to them in his environment. But why did he learn these words, and in the order that he learned them? In the next chapter, we consider the relationship between lexical acquisition and the rich linguistic environment of a young child's first years.

Chapter 5

Environmental Contributions to Word Learning

Children's early language learning is sometimes described as "effortless", and to adults witnessing the seemingly autonomous birth and growth of language it may indeed appear so. But a better adjective might be "remarkable" when one accounts for the numerous challenges that young learners face in acquiring their first language. Unlike adults learning a second language, young children do not possess the full array of concepts that are the subject of everyday speech. Word learning requires more than assigning labels to a preexisting library of meanings, as the conceptual framework for understanding the world is just one of the actively developing "strands" described in (Snow, 1988). Even the labels themselves – the words – are not necessarily directly accessible. Children's exposure to language is primarily through speech, and unlike text there are no "spaces" marking word boundaries. As Peters (1983) discusses, although the units of speech are words, children do not necessarily partition the speech stream into their final adult word forms. Even assuming the words and the concepts are available to the child, the mapping between them must be learned. In discussing how children learn the names of simple ideas and substances, Locke (2008) famously claimed that adults ordinarily show children "...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." In fact, such ostensive naming by adults is the exception rather than the rule (Bloom, 2000).

Despite these many challenges, children do come equipped with a range of sensitivities and skills for language acquisition. Elizabeth Spelke and her colleagues argue that children come into the world equipped with systems of *core knowledge* about objects, agents, number, geometry as well as social knowledge (Spelke, 1994; Spelke and Kinzler, 2007). Such systems of core knowledge may provide a necessary substrate for early learning, including language acquisition. Children are also sensitive to statistical regularities in the speech they hear, which can help in segmenting words (Saffran et al., 1996). Another skill children bring to bear, of particular relevance to word learning, is the ability to infer the referential intent of others. In the case of learning names for objects, a child must associate the name to what the *speaker* is referring to, even if that is not the child's focus of attention when the name is uttered (Baldwin, 1991).

These last two examples highlight the interplay between the child's pattern detection and inference mechanisms and the learnable structures provided by the environment. We use the term "environment" in a broad sense, including the wide variety of experiences, routines, people and things that make up the child's world. As Paul Bloom (2000, p. 90) says, "People cannot learn words unless they are exposed to them. We can explain much of the character of children's vocabularies in terms of this banal fact" and as such, characterizing the learning environment is crucial in understanding early word learning.

5.1 The Learning Environment

The wide range of conditions for early learning, which can vary greatly across cultures as well as across families, can make characterizing the learning environment a daunting task. A recent study on language development in the extreme conditions of very limited social and linguistic interaction is described in (Windsor et al., 2007). Through the Bucharest Early Intervention Project (Zeanah et al., 2003), children in Romanian orphanages – historically noted for their sad conditions of extreme social and physical deprivation – had been tracked and compared to children that had been moved into foster care. In (Windsor et al., 2007), the authors assessed language measures for ten 30-month old children in orphanages, comparing them to ten age-matched children recently moved to foster care, ten children in foster care for more than a year, and ten children raised in their biological families. The overall results showed that children that had been in foster care for more than a year, or that had never been institutionalized, had (mostly) comparable linguistic performance but were far beyond their peers still in orphanages or recently moved to foster care. Notably, 4 children from these latter two groups did not produce more than *one* intelligible word at 30 months of age, although the sample size is insufficient for drawing broader conclusions about the cohort.

Further discussion on the "origins" of language in children in atypical situations, including feral children that have had essentially no human contact in their early years, is presented in (Comrie, 2000). In this speculative paper, the author considers the minimum that the environment must provide for language to develop. The author suggests that access to a lexicon, which may trigger some insight into the arbitrary relation between linguistic form and meaning, may be the minimal environmental requirement. In addition, access to a community of potential speakers is also critical. While many may agree that the origin of language in children is both biological and cultural, the author offers interesting ideas on how to delineate between these two powerful forces.

But what about the environmental variability in the rest of the cases – the overwhelming majority of "normal" situations where children are part of a linguistic community and successfully learn their native language? In the case of word learning, strong evidence for the positive link between the total amount of maternal speech and children's vocabulary size was provided by Hart and Risley (1995). But what is the relationship between the words children hear and the words that they learn? Huttenlocher et al. (1991) investigated this question at a time when the prevailing assumption was that learning capacity accounted for much of children's individual differences in vocabulary. However, they found a strong positive connection between the frequency of words in caregiver speech and vocabulary growth rate and order of word acquisition. Goodman et al. (2008) conducted further work along these lines, but with a larger sample of words. They found the same overall effect: more frequent words in caregiver speech are learned earlier, but crucially, only when part of speech is taken into account. However, their analysis was based on estimates of "parental input" and child vocabulary from completely separate corpora. That is, the vocabulary under investigation was derived from MacArthur-Bates CDI measures and the input was derived from caregiver speech corpora available in CHILDES (MacWhinney, 2000).

Exposure to caregiver speech affects more than just the words that are learned. In recent work, Hurtado et al. (2008) showed that it also positively impacts children's speech processing efficiency. Children exposed to more caregiver speech at 18 months knew more words and were faster at word recognition at 24 months. One of the interesting results of this study was the substantial overlap in the effect of maternal speech input on these two outcomes, suggesting that increased processing efficiency supports faster lexical learning, but also that greater lexical knowledge contributed to faster processing efficiency. To use Snow's analogy, these findings suggest that the developmental "strands" of speech processing skill and lexical knowledge are both entangled and mutually supportive.

Another crack in the chicken-and-egg problem of learning words from speech, where the speech cannot be parsed into words because the words are unknown, was studied in (Brent and Siskind, 2001). The authors found that isolated words were a reliable feature of speech to 9-15 month old children and that the frequency of their use in isolation better predicted their acquisition than overall frequency. In light of the (Hurtado et al., 2008) study, the picture that emerges is one where the environment provides a variety of footholds for the young learner, some of which may be easily accessible and some which may become more accessible with development.

5.1.1 Analysis overview

This work focuses on lexical acquisition in the Speechome Corpus, and the role that environmental factors play in one child's early word learning. The basic analysis strategy is to quantify variables of interest for words in the child's "input", and to regress these variables against the child's age of acquisition for those words. We begin by characterizing the linguistic environment in terms of overall word exposure and the timing of how words are experienced. Although more work could be done with linguistic aspects of the input, our interest is in further filling out the picture of the child's environment. Toward this end, we study the spatial use of language, finding that the spatial distribution of a word's use in the house is a strong predictor of when it is learned.

5.2 Word usage frequency

The child's overall exposure to a word prior to the word birth can be directly quantified in the Speechome Corpus. In principle, we wish to find the number of uses of a word in caregiver speech (or non-child speech) in circumstances where the child could be considered "exposed" to the word. In practice, this can be accomplished by counting occurrences of the word in "child-available speech" prior to the word birth. By considering utterances prior to the word birth, we can safely assume that none of the word occurrences were produced by the child. Furthermore, since we are not concerned with speaker identity in this analysis, we can also include occurrences in multispeaker utterances.

Our emphasis is on child-available speech (CAS) since it can be objectively identified through a combination of "where-is-baby" annotations and utterance location, but researchers more often focus on child-directed speech (CDS). An important question is how child-directed and child-available speech differ in terms of their impact on word learning. Weisleder and Fernald (under review) provide further evidence for the importance of child-directed speech in lexical acquisition, finding that overheard speech by itself contributes little. However, in their work the superset of "meaningful speech", which includes both child-directed and overheard speech, preserves the relationship to vocabulary growth and is most directly comparable to CAS (Weisleder, 2011, personal communication). In addition, recent work developing methods for identifying CDS in the Speechome data (Vosoughi and Roy, 2012) revealed that roughly 72% of CAS is actually child-directed, when balancing the data by speaker and by month of recording. In fact, the percentage of CAS that is child-directed is actually higher (roughly 83%) when a larger random sample that is not balanced by speaker is used (Vosoughi, 2012, personal communication). Thus CAS is actually well aligned with CDS and is a reasonable source of input speech to study.

The number of occurrences of a word prior to its acquisition will vary as a function of both word frequency and the AoA itself. For words with the same frequency, those learned later will tend to have a higher occurrence count, following the simple relation that $E\,[\text{total occurrences}] = \text{AoA in days} \times E\,[\text{num occurrences per day}]$. Of course, the actual number of occurrences in a day varies, and depends on the probability of choosing the word and the total amount of speech in a day. But we are interested in capturing the child's experience of one word relative to another, and the average number of occurrences per day serves as a useful measure of exposure. Therefore, word frequency is calculated as

$$\text{freq}_w = \frac{\text{count}(w) \text{ prior to AoA}}{\text{num days prior to AoA}}$$

The caregiver word frequencies are highly positively skewed, with just a few very frequent words and many infrequent words as shown on the left in figure 5-1. This is in line with the general observation of Zipfian distributed word frequencies noted in section 3.3. Following (Huttenlocher et al., 1991), we transform the usage frequencies by the logarithm. This is appropriate for positive data, and in particular data that spans several orders of magnitude as ours does. The log-transformed frequency histogram appears roughly symmetric, as shown on the right in figure 5-1. We use $\log_{10}$ for ease of interpretation – a change in 1 unit corresponds to a factor of 10 in word usage frequency.

For the 669 word births, the relationship between log-usage frequency and word birth date is shown in figure 5-2. The relationship shows that words used more frequently in caregiver speech are learned earlier by the child. Both slope and intercept terms in the model are significant, as seen in table 5.1. The slope coefficient of -0.46763 indicates that using a word more by a factor of 10 shifts the word birth earlier by about half a month.

It may not be surprising that a word's frequency in the child's input relates to when it is learned. After all, the young child will not learn words that he is never exposed to. On the

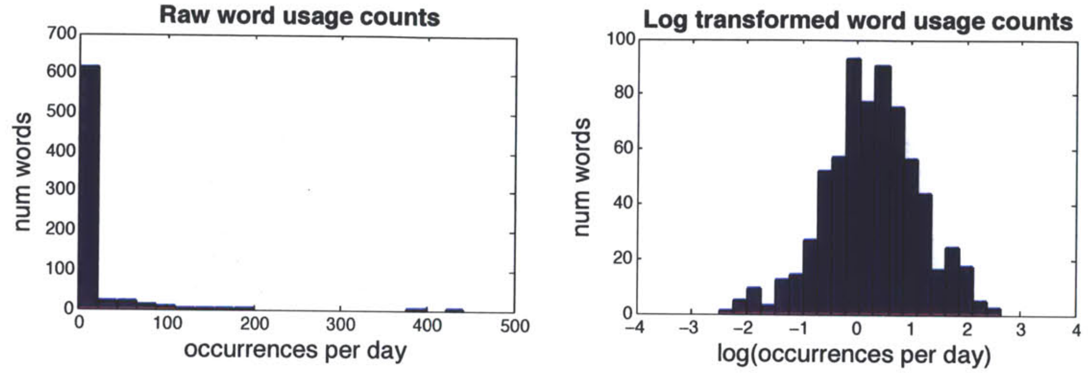


Figure 5-1: Histogram of caregiver usage frequencies for the raw and log-transformed occurrence frequencies. This is for the set of words the child learned.

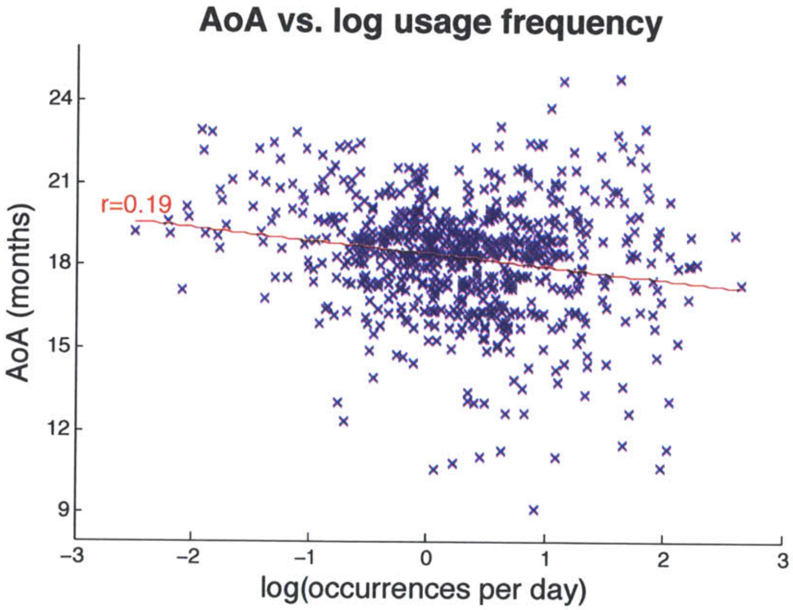


Figure 5-2: AoA for each word in child's vocabulary against the log of caregiver usage frequency, and regression line showing the effect of frequency on AoA. Caregiver word usage and AoA are negatively correlated, meaning that words used more frequently are learned earlier with a correlation coefficient of r=.19.

	Estimate	SE	tStat	pValue
(Intercept)	18.405	0.085008	216.5	0
log_usage	-0.46763	0.095198	-4.9122	1.1304e-06

Table 5.1: Regression model for log usage frequency

other hand, every exposure to a word is an opportunity to make some link between sound and meaning, and whether learning is driven by exposure statistics and the accumulation of evidence (Yu and Smith, 2007; Smith and Yu, 2008) or by forming and testing hypotheses about a word's meaning (Medina et al., 2011), the number of exposures should matter. But while more frequent words tend to be learned earlier, the frequency predictor has some additional subtlety. In our earlier work (Roy et al., 2009; Vosoughi et al., 2010) we found that the effect of frequency varied considerably by word class. Nouns showed a stronger frequency effect and are generally learned earlier, although their overall frequency is less than closed-class words, which are learned later and show a weaker frequency effect. The variation across different word classes was also noted in (Goodman et al., 2008).

5.3 Word recurrence

Word usage frequency captures the child's overall exposure rate to words in caregiver speech at the timescale of a day. However, speech occurs at the timescale of seconds, and words with similar daily frequencies can have very different usage patterns at shorter timescales. We calculate word *recurrence* to measure the degree of repetition of a word in a one-minute period.

The recurrence of a word w is calculated as follows. For a sequence of contiguous, non-overlapping time intervals t_i, count the number n_i of caregiver uses of w that fall in each interval. The time intervals span from the child's 9 month age point up to the age of acquisition of w. This yields a set of time intervals and counts $\{(t_i, n_i)\}$. In calculating recurrence, we only consider the subset in which the word occurs and take the average over the counts. More precisely,

$$\text{recur}_w = \frac{\sum_i n_i}{\sum_i \mathbb{1}_{[n_i > 0]}}$$

where $\mathbb{1}_{[\cdot]}$ is the indicator function with value 1 if the condition is true, 0 otherwise. Figure 5-3 illustrates the recurrence calculation and a situation where two words have the same frequency but difference recurrence values.

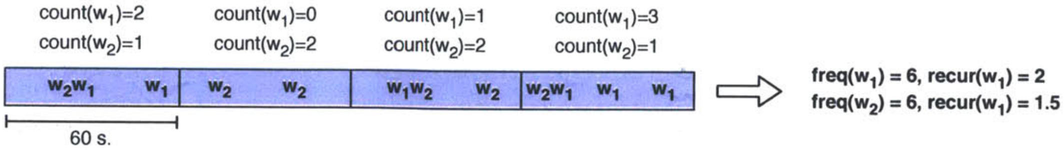


Figure 5-3: Illustration of the recurrence calculation for words w_1 and w_2. Note that they could both have the same frequency and different recurrence values, as w_1 only occurs in three intervals while w_2 occurs in four intervals. In practice, the number of intervals generally differs for each word, since the rightmost interval depends on the AoA for the word under consideration.

There are several reasons that caregiver word recurrence may be an interesting variable to consider in analyzing the child's early word learning. Recurrence reflects some sensitivity to working memory and attentional limitations – a word heard repeatedly in a short period of time may be more salient than the same number of exposures spread out in time. This was an important observation in the development of the CELL model, a model of cross-modal word learning (Roy, 1999; Roy and Pentland, 2002). In that model, which learned word meanings from audio-visual input, an exhaustive search for recurring acoustic patterns identified candidate "audio-visual prototypes" for word-meaning pairs. Limiting the exhaustive pattern search to the contents of short-term memory was a requirement for computational tractability. But Roy and Pentland (2002) also point out that target words were usually repeated multiple times within short time windows in caregiver speech, a finding supported by several different datasets the authors examined.

If the learner's working memory limitations determine which words can be learned, then a word's recurrence value should be predictive of when it is learned. But another reason to consider recurrence has less to do with the learner's limitations and more to do with how language use relates to activities. A word repeated frequently in a short period of time may play an important role in an activity that is currently taking place. For example, in playing with a ball, the word "ball" may be used frequently over the duration of the activity, and then used little throughout the rest of the day. Recurrence may be a good measure of whether a word is salient in particular contexts. It need not be salient in *all* contexts to have a high recurrence, but if is salient in some situations and is repeated during those situations it can have a high recurrence while having a low frequency.

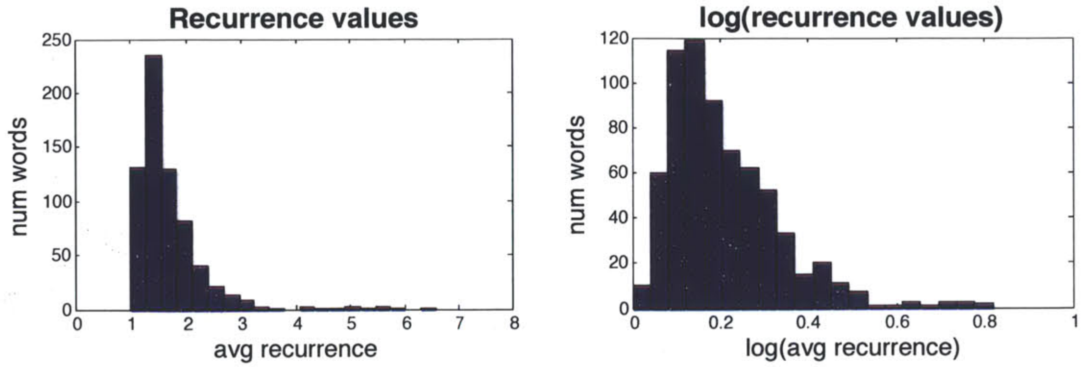


Figure 5-4: Histogram of raw and log-transformed recurrence values. This is for the set of words the child learned.

It is also possible that recurrence captures caregiver sensitivities to the child's interests and current lexical knowledge. When caregivers use a word in a more recurrent fashion, it may really be an effect on the caregiver: they are tuning in and responding to the child's current state of knowledge. That is, if the child seems to understand a word even though he hasn't yet produced it, caregivers may use the word at an increased rate (although if this were the case, it should apply to frequency as well.) This idea highlights the question of causality in shaping the learning environment, and the mutual influence of both caregiver and child on one another.

Vosoughi (2010) first explored recurrence in the Speechome data, considering a range of temporal window sizes and finding that a duration of 51 seconds was most predictive of AoA. That analysis, based on an earlier corpus and word birth list, was extended and reported in (Vosoughi et al., 2010). However, in our current analysis we examine recurrence in `corpus12` and do not optimize the window size to maximize correlation, but instead round up to a 60 second window for simplicity. We also found little difference between the 51 second and 60 second window with this data.

Figure 5-4 shows the distribution of the word recurrence across all words in the child's vocabulary. The histogram on the left is the raw recurrence, while the histogram on the right is the log-transformed recurrence (base 10). As with frequency, the distribution is positively skewed, and log-transforming the data better conditions it for analysis. The relationship between log-recurrence and AoA for each word is shown in figure 5-5, along with

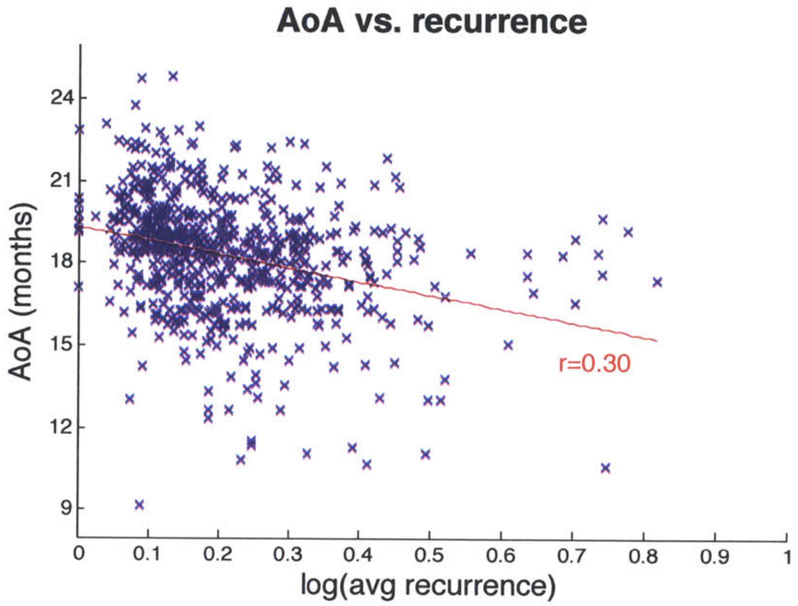


Figure 5-5: AoA for each word in the child's vocabulary against the log of caregiver recurrence value, and regression line showing the effect of recurrence on AoA. Caregiver recurrence and AoA are negatively correlated, meaning that words with higher recurrence are learned earlier with a correlation coefficient of r=.30.

	Estimate	SE	tStat	pValue
(Intercept)	19.304	0.14995	128.74	0
recur	-2.1538	0.26582	-8.1024	2.5178e-15

Table 5.2: Regression model for log average recurrence

the regression line. The correlation coefficient $r = 0.30$ demonstrates that log-recurrence is a better predictor of AoA than frequency, which has a smaller correlation coefficient. Table 5.2 summarizes the linear model. The effect of log-recurrence on AoA is negative, implying that AoA shifts earlier as word recurrence increases.

5.4 Spatial factors

Recurrence and frequency capture aspects of the child's linguistic exposure, but the child's experience with language takes place in a rich, multimodal domain. The stream of words that serve as "input" occur in the child's home, and the spatial context of where words are heard may be an important variable to consider. One of the main reasons for capturing video in the Speechome Corpus is to study language in context, and spatial context may be both relevant to word learning and directly obtainable from video.

This section is based on work by Matthew Miller (Miller, 2011), who developed techniques for automatically identifying coherent regions from raw video and extracting activity patterns over these regions. He found that activity patterns were predictive of word births in the Speechome data. Related work by George Shaw (Shaw, 2011) on the Speechome data used custom person tracking algorithms to follow people as they moved about the house. Shaw characterized movement in the house and also found a strong link between where words are used and when they are learned.

5.4.1 Representing video

The first challenge in studying the spatial context of language use is to extract a meaningful representation of spatial activity from video. Given the amount of video in the Speechome Corpus, the lower resolution WINKs were processed. First, a simple background subtraction procedure was applied, in which a pixel is considered *foreground* if the difference between its intensity and the average intensity for that pixel is greater than a threshold. The resultant sequence of binary frames are then processed to identify coherent regions. A coherent region is one where pixels in the region tend to covary with one another. An affinity matrix A was constructed with A_{ij} representing the chance that both pixels i and j are labeled as foreground if either is labeled as foreground. For practical reasons, affinity was not calculated for all pairs of pixels, but only for those within a certain radius of each other. A is then decomposed into clusters such that all clusters are below a distortion threshold (inversely related to the affinity of pixels in the cluster.) Clusters above threshold are repeatedly split and all cluster boundaries are updated until the distortion criterion is met. The distortion threshold provides a parameter to roughly control the number of regions discovered. The highest resolution partitioning yielded 487 regions over 9 of the 11 cameras in the house, reproduced from (Miller, 2011) in figure 5-6.

Background subtraction, which served as a first step in processing the video for identifying

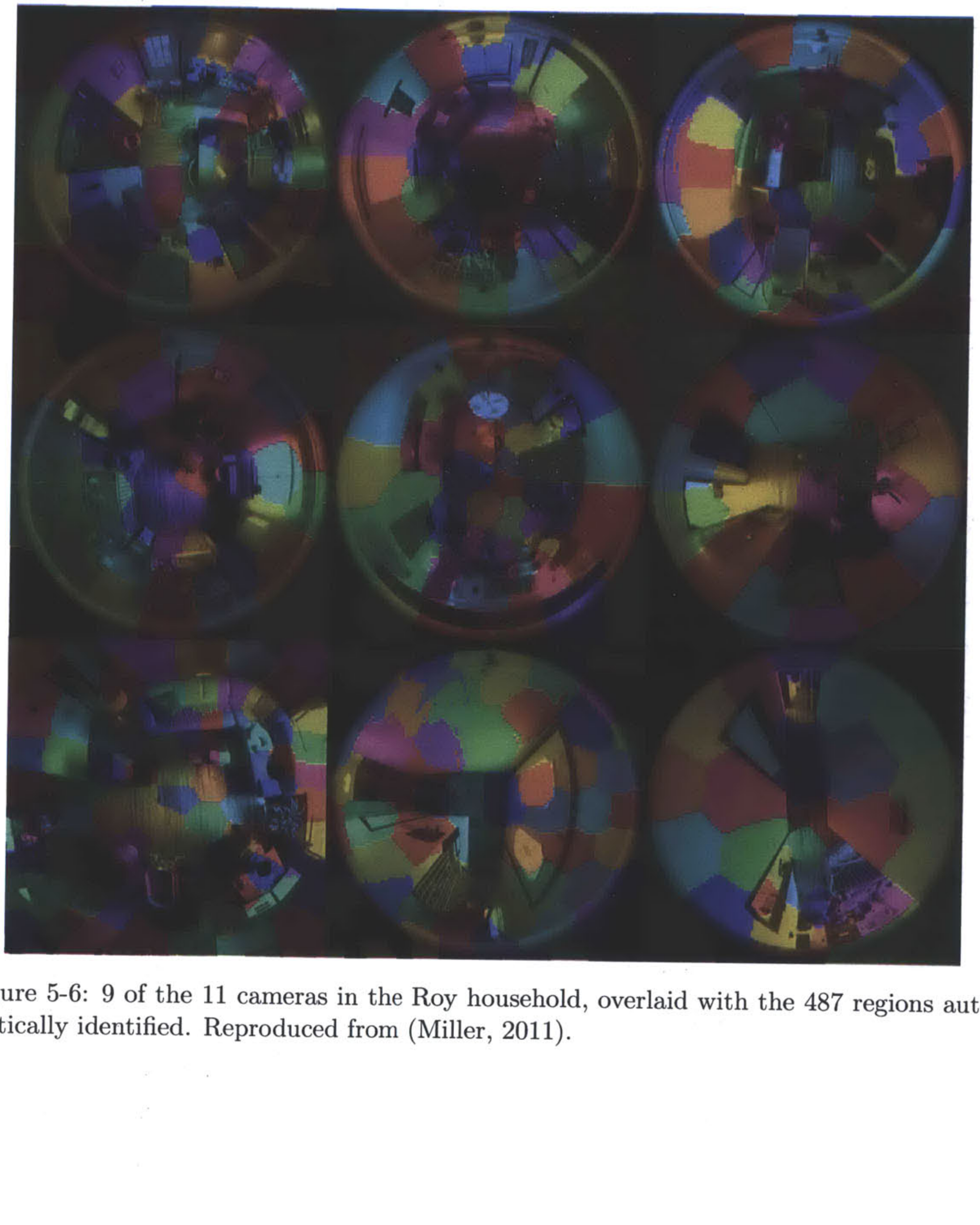


Figure 5-6: 9 of the 11 cameras in the Roy household, overlaid with the 487 regions automatically identified. Reproduced from (Miller, 2011).

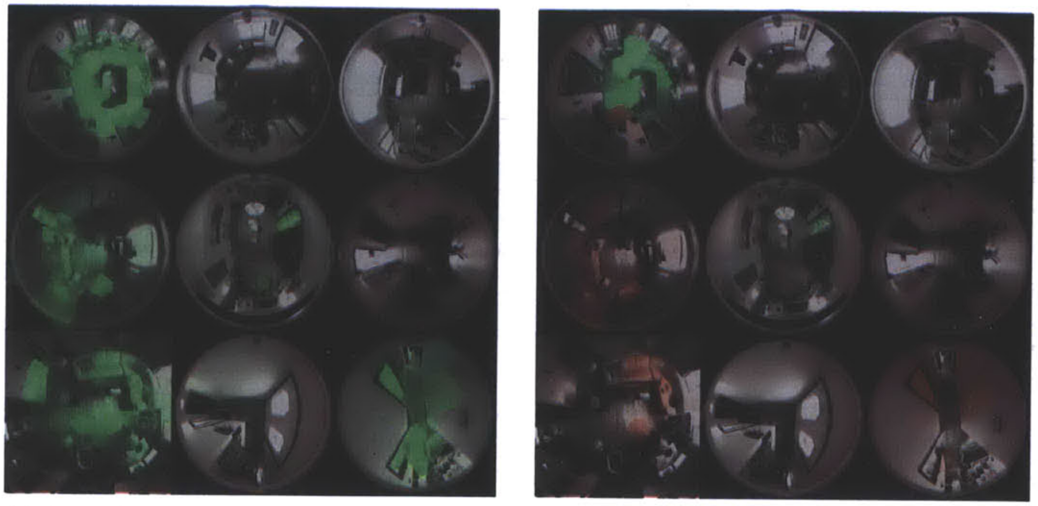
regions, is often also a first step in algorithms for tracking motion in video. However, Miller argues that more sophisticated tracking techniques may introduce errors in seeking precise spatiotemporal trajectories, but for capturing general spatial activity foreground pixels can be used directly. While a single foreground pixel may not reflect human motion, a sufficient number in aggregate can be a reliable indicator of human activity. Miller labels a region as active for a timespan if a minimum number of pixels are foreground in the relevant frames. Alternatively, activity can be quantified by counting the foreground pixels in the region for the timespan.

5.4.2 Characterizing a word's spatial usage patterns

The low dimensional representation of video and spatial activity developed in (Miller, 2011) offers a new view of the Speechome data. The activity distribution derived from all video across all regions shows how activity is distributed throughout the home. Considering a subset of this data, for example, just the activity around uses of a particular word can be revealing. Figure 5-7(a) shows the average distribution of all adult speech, and figure 5-7(b) shows spatial activity around uses of the word "coffee" relative to this average. These were generated by computing the region activity for all video frames in a 10-second time window centered on an utterance, and averaging over the appropriate set of utterances.

The word "coffee" has usage patterns distinct from overall adult language use located where one might expect: near the coffee maker in the kitchen. It is not surprising that word use relates to place – different parts of the house serve different functions, and the concomitant human behavior and language are naturally linked. Nevertheless, a data driven exploration opens the door to finding unexpected patterns in everyday behavior.

This leads to the question of whether a word's spatial usage characteristics might relate to the child's lexical uptake. To investigate such a relationship, we'd like to find a meaningful, quantifiable property of a word's spatial usage patterns and relate that to the age of acquisition for that word. This follows along the same lines as using frequency and recurrence as measures of caregiver input speech. Motivated by the idea that a word may be more salient



(a) Spatial activity around overall adult language use. This constitutes a "background" activity distribution.

(b) Relative spatial activity around caregiver utterances of the word "coffee" as compared to the background, with more activity in green regions and less in red regions.

Figure 5-7: Spatial activity calculated over video frames falling within ±5 seconds of particular utterances, focusing either on all adult utterances (left) or just those utterances containing the word "coffee" (right), with most activity near the coffee maker in the kitchen. Reproduced from (Miller, 2011).

if it is used in more predictable ways, and in ways that are distinct from the sea of all language, we quantify how a word's spatial distribution deviates from the overall distribution. The Kullback-Leibler divergence (KL-divergence) measure, also known as relative entropy, captures how two distributions P and Q differ, defined as $D(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$. KL-divergence is measured in bits when using $\log_2$, or "nats" under the natural logarithm. It is an information theoretic measure that captures the number of additional bits (or nats) required to encode P with an encoding scheme based on Q (Cover and Thomas, 2006). While not a true distance metric, it has the property that $D(P||Q) \geq 0$, with equality if and only if $P = Q$. To make this more concrete, the discrete probability distribution over the 487 regions for the word "coffee" is quite different from the background distribution Q, as indicated by the amount of red and green in figure 5-7(b). Consequently, $D(P_{\text{coffee}}||Q)$ will be large. On the other hand, a word such as "the" is very similar to the overall background distribution, and $D(P_{\text{the}}||Q)$ will be close to 0.

The fact that P and Q are distributions over space is *not* captured by the KL-divergence

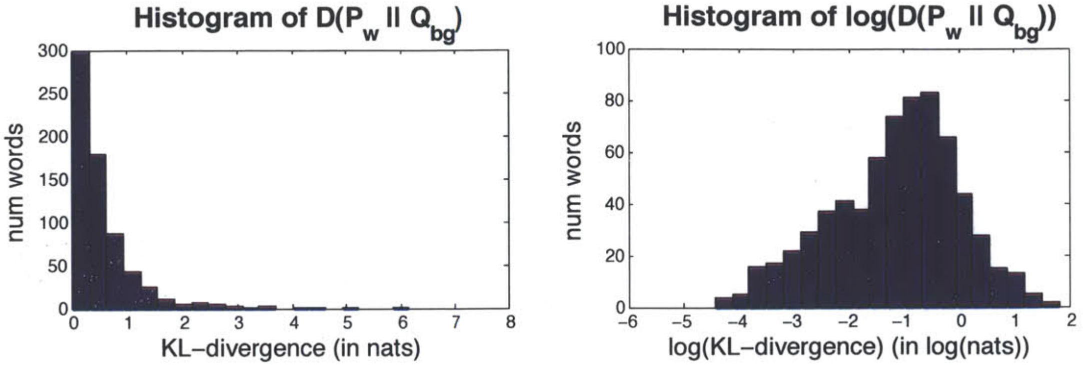


Figure 5-8: Histogram of the KL-divergence between a word's spatial distribution and the background spatial distribution of overall language use, for the words learned by the child. The raw KL-divergence distribution is skewed, which is helped by log-transforming these values as shown on the right.

measure. Other statistical measures, such as Ripley's K and Moran's I respect the spatial structure of P and Q, and these are explored further in (Shaw, 2011). Here, we simply acknowledge that large $D(P||Q)$ indicates that P differs from Q, although it does not necessarily imply that P is spatially more coherent. Shaw (2011) illustrates cases where a spatial statistic such as Ripley's K may be preferable to a more general measure such as KL-divergence.

5.4.3 Relating spatial activity to word learning

Figure 5-8 shows a histogram of the KL-divergence relative to the background for each word in the child's vocabulary. As mentioned above, the spatial activity for a single utterance is calculated over all frames falling within ±5 seconds of the utterance. The background distribution is the average spatial activity over all non-multispeaker, caregiver utterances in the corpus. The activity distribution for each word is calculated over all non-multispeaker, caregiver utterances containing the target word up to the word birth date. As can be seen, the distribution of KL-divergence values is skewed, and applying a log-transform better conditions the data for analysis.

Figure 5-9 shows the result of regressing AoA on log KL-divergence. The regression line indicates a tendency for words with higher log KL-divergence (and thus higher KL-divergence)

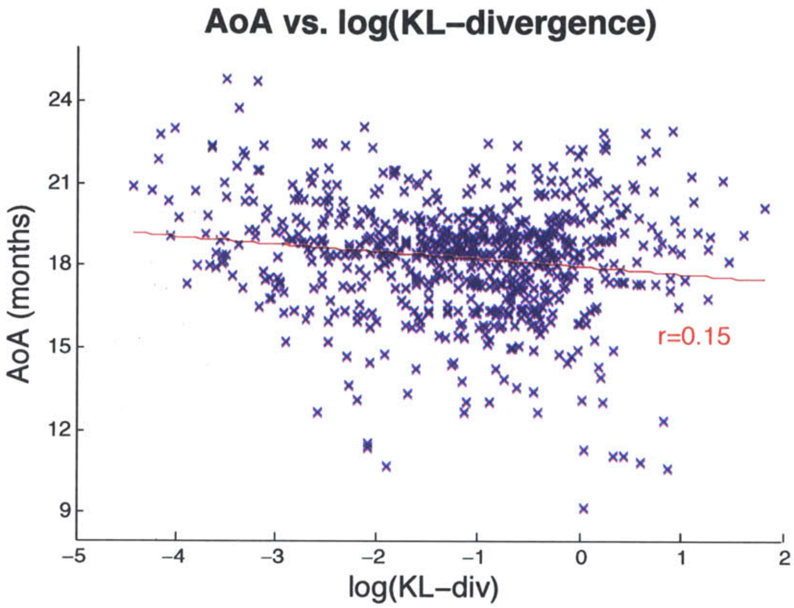


Figure 5-9: Relationship between the log-transformed KL-divergence and AoA. As KL-divergence increases, AoA shifts earlier, indicating that words with more distinct spatial usage patterns from overall language use tend to be learned earlier.

to be learned earlier. However, the correlation with AoA is weaker than that of frequency. The finding is interesting nonetheless: words with distinct spatial distributions (relative to the background) tend to be learned earlier.

Before going further, there is an important confound that must be addressed. The spatial distributions used are really the empirical distributions $\widehat{P}$ and $\widehat{Q}$, obtained via a finite set of measurements as estimates of the true P and Q. The number of measurements is essentially the number of utterances used in constructing the distribution, and the number of utterances available for different words varies considerably.[1] We should expect that as the number of samples n increases the empirical distributions will approach the true distributions, but what effect will this have on the KL-divergence measure? Intuitively, distributions built up from fewer samples will be "sparser", since the activity for any single utterance will likely only occupy a small number of regions. As the number of samples increases, the distribution will fill out and begin to look more like the smooth background, leading to a reduced KL-

[1] Although the number of video frames and the amount of activity for different utterances may vary, we ignore this in considering the sampling process. Each "sample" is a snapshot of the spatial activity corresponding to an utterance.

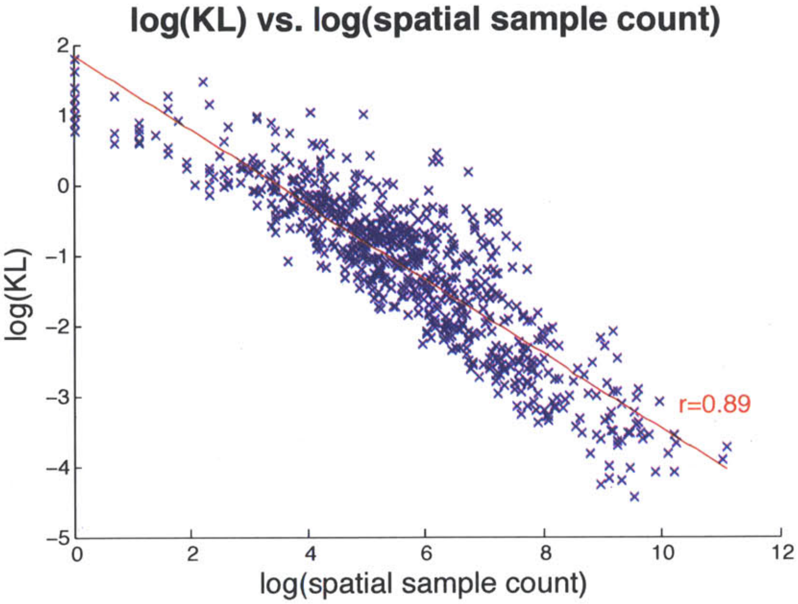


Figure 5-10: Relationship between the log-transformed KL-divergence and log sample count. This shows that the KL-divergence measure for a word is highly dependent on the number of samples used in obtaining the KL-divergence, a confound for our AoA regression analyses.

divergence as n increases. More technically, the KL-divergence based on $\widehat{P}$ and $\widehat{Q}$ is a biased estimator of $D(P||Q)$. This relationship is borne out empirically, as shown by the strong linear relationship between log KL and log count in figure 5-10. The mathematical underpinnings of this relationship are derived and further explored in appendix A.

What effect does this have on the regression with AoA? First, as count increases KL-divergence decreases, and if the relationship between KL-divergence and AoA holds we should expect a later word birth. On the other hand, a higher word count is also associated with a higher word frequency which we have found to correlate with an earlier word birth. KL-divergence and frequency seem to be competing with each other, linked through word count. Returning to figure 5-10 suggests that the dependence on sample count may be removed by taking the log KL-divergence residuals. The count-corrected log KL divergence proves to be strongly correlated with AoA, as shown in figure 5-11. Miller (2011) obtains essentially the same result, but using an earlier (and smaller) list of word births and an earlier version of the corpus with fewer transcribed utterances. In addition, his treatment of multispeaker utterances and other transcript processing steps differed.

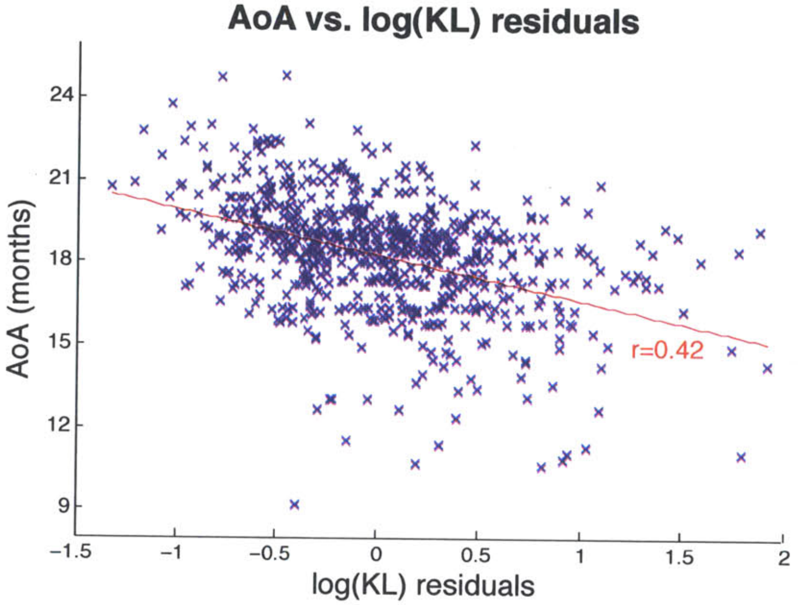


Figure 5-11: Relationship between the log-transformed KL-divergence *residuals* (as predicted by log sample count) and AoA.

5.4.4 Discussion

The key result in this section is that the log KL-divergence residual for a word in caregiver speech is negatively correlated with its age of acquisition. Controlling for word frequency, a word with a spatial distribution that is more distinct from the average tends to be acquired earlier, and the effect is also stronger than the linguistic predictors of frequency and recurrence.

This result can be interpreted in several ways. One interpretation, and perhaps the most direct, is that words that are more "spatially unique" are somehow easier to learn. If the child has certain overall expectations about where words are used, then those words that deviate from these expectations may be more salient. One way for a word to deviate from what is expected is if it tends to be used only in certain places. Although KL-divergence itself is not a spatial measure, such words will tend to have a high KL-divergence from the overall distribution. One can see how words that are spatially localized might be more salient to the child. For example, a particular chair may be strongly linked to caregiver use of a certain word, effectively providing a restricted, manageable context for the word's use. However, as noted above, a word's distribution may also deviate from the average without being spatially localized, an issue considered in more depth in Shaw (2011). If such non-localized but spatially distinct words are salient to the child, it may be because they deviate from the expectations set up by overall spatial word use characteristics. Or such words may just be less salient, and KL-divergence is really serving as a proxy for spatial localization.[2]

Another interpretation, which assumes that a word's spatial uniqueness primarily derives from being spatially localized, is that such words are tied to activities that are also localized. Episodes of focused child-caregiver interaction when the child is engaged in certain activities may be the real contributor to word learning, but these activities may also be spatially grounded. For example, reading books with a caregiver might usually happen in a certain chair, so the words used during book-reading will be tied to the location although their uptake is driven by the activity. The basic idea here is that the spatial uniqueness of a word is a proxy for the word's salience in activities that make up the child's early experience. This idea is further explored in the next chapter.

5.5 Conclusion

Early word learning is a remarkable feat, achieved when young children's powerful learning mechanisms, biases and sensitivities meet the rich, structured environment of their first years of life. Children learn words through experience with language, but this experience is part of a broader social and physical context. By characterizing aspects of a child's early experience with language, factors that contribute to early word learning can be identified and compared. Studying the environmental contributions to word learning can shed light on the learning mechanisms and biases that children bring to the task.

One of the simplest measures of linguistic experience is the frequency of word use in caregiver speech, which we find to be predictive of when words are learned. More frequent words

[2] The latter suggestion, that non-localized but spatially distinct words are less salient, seems more likely, and some supporting evidence for this is presented in (Shaw, 2011).

tend to be learned earlier, in accordance with prior research. With a detailed record of all caregiver speech prior to a word's acquisition, other measures such as recurrence can also be calculated. Recurrence is more predictive than frequency, indicating that it may better capture what is salient in the child's experience. But word use is structured along many other dimensions. The spatial aspect of how words are used proves to be strongly predictive of when words are learned – words that are used in more "spatially unique" ways tend to be learned earlier. One interpretation of this finding is that spatially unique words may be strongly tied to particular activities, which may be an underlying source of their "learnability". In the next chapter, we attempt to uncover the activity structure of daily life, and link that to the child's early word learning.

Chapter 6

Language Use in Activity Contexts

Children do not learn language in a vacuum – their early learning environment is rich and multifaceted. In the previous chapter, several environmental factors were characterized and shown to contribute significantly to early word learning. Both linguistic and nonlinguistic factors were considered, with the perhaps surprising result that the spatial aspect of a word's usage patterns was strongly predictive of when it was learned.

For learning to take place, there must be some consistent structure that is accessible to the learner, as well as learning biases to help guide them toward extracting appropriate patterns from the experiential data to which they are exposed. Much attention has been paid to the biases inherent in learning syntax, largely motivated by Chomsky's "poverty of the stimulus" argument (Chomsky, 1965, 2005). If children's exposure to language does not provide sufficient data for inferring grammatical structure, then those structures, or their antecedents, must already be present at birth. Such an innate language faculty is sometimes referred to as a "language acquisition device" (LAD).

A contrasting view, which does not push all the work of language acquisition onto a specialized, innate mechanism is presented in (Bruner, 1985). While Bruner acknowledges the role of innate human faculties in language learning, he focuses on the social structures that support communication and guide children toward acquiring their native language. These social structures are part of the child's natural context and early learning environment, and Bruner makes the case for studying language learning in this setting. Bruner (1983, p. 9) 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." This "clutter" was the setting for the study reported in (Bruner, 1983), which focuses on two children in their homes and their everyday interactions with their parents. Bruner found that games and routines provided an important scaffold into language – supporting communication with the prelinguistic infant, serving as a structure for meaningful word use, and providing a model for discourse and conversational exchange. He referred to this important foundation as the *language acquisition support system* (LASS), which centered on the notion of a *format*. A format is a stable structure in which both child and caregivers participate, a "...a rule-bound microcosm in which the adult and child do things to and with each other." (Bruner, 1985).

6.1 Formats and activity contexts

The primary format studied in (Bruner, 1983) was "peek-a-boo", a game involving caregiver and child that began before the child's first words and grew in sophistication with the child's development. This game was enjoyed by mother and child over the course of the study, with the mother continually demanding more expressive language and behavior from the child. Eventually, the child began to take on the role of the primary actor in the game. Overall, formats are rich patterns of activity, with social roles as well as a deep structure the can manifest in a variety of different ways. For example, peek-a-boo may be played by covering one's face or by hiding behind a chair; what is important is the disappearance/reappearance sequence. The restricted world of peek-a-boo helps to constrain word meanings and provides a consistent framework for linking speech acts to the state of the game and the actions of the other participant. The contingence of each participant's action on a prior action or response by the other helps to model discourse structure. But above all, formats capture patterns of communicative, interactive behavior between participants.

If we take the idea of the LASS seriously, there should be many formats in addition to games like peek-a-boo. What other recurrent, structured activities make up a young child's daily experience? One of the key challenges to developing and testing a theory of how structured social interaction might lead a prelinguistic child into language is simply identifying the formats. This is a prerequisite for a characterization of their deep structure, social roles, and other aspects. From the clutter of (recorded) daily life, what are the formats, when do they occur, and moreover, in the face of large-scale naturalistic recordings, how can we operationalize their detection and analysis?

In this work, we propose to identify *activity contexts* as the first step toward a large-scale analysis of formats in the Speechome Corpus. Rather than modeling deep structure, activity contexts effectively label *what is happening* at the temporal granularity of minutes. Mealtime, story time, playing with toys, and of course, peek-a-boo, are all possible activity contexts. As a label for a segment of recorded data, activity contexts indicate that a particular pattern of behavior is exhibited in the data. These labels provide a means for partitioning the corpus into contextually coherent subsets of data for analysis. Rather than considering overall language use, the child's linguistic experience can be considered in context. For example, language used during "mealtime" episodes may have certain consistent elements that are lost when mixed with all language in the corpus.

6.1.1 Could activity contexts shed light on word learning?

So far, we have appealed to ideas such as Bruner's *language acquisition support system* to motivate the use of activity contexts in studying early word learning. However, another hint at their potential value comes from a reconsideration of environmental factors discussed in the previous chapter. Consider word frequency and word recurrence from the perspective of the child. Word frequency is a relatively global measure, summarizing all caregiver uses of the word from 9 months to the word birth date. On the other hand, recurrence is a local measure: given that the word occurred, what is its occurrence rate over the timespan of a minute? A word like "mango" may have a high recurrence value even if its overall frequency is low, since when a mango is relevant to the situation it may be a subject of both attention and conversational discourse and used frequently within a short period of time. Words with a high recurrence value may be aligned with focused activities or situations, and notably, word recurrence is a better predictor of AoA than word frequency. As Bruner (1985, p. 36) says, "If there is a Language Acquisition Device, the input to it is not a shower of spoken language but a highly interactive affair shaped, as we have already noted, by some sort of an adult Language Acquisition Support System." Word frequency effectively ranks a word in the "shower of spoken language", while word recurrence may be capturing something of the word's salience to the child's experience.

The same may be said for the spatial activity feature described above, which is perhaps even more easily interpretable as an activity context indicator. Words that have a spatial distribution that is particularly distinct from overall language use may arise due to an activity tied to a particular location, such as diaper change, book reading, and so on. As Miller (2011) pointed out, spatial activity patterns may be capturing coherent behaviors and activities that are relevant to word learning.

6.2 Manually annotated activities

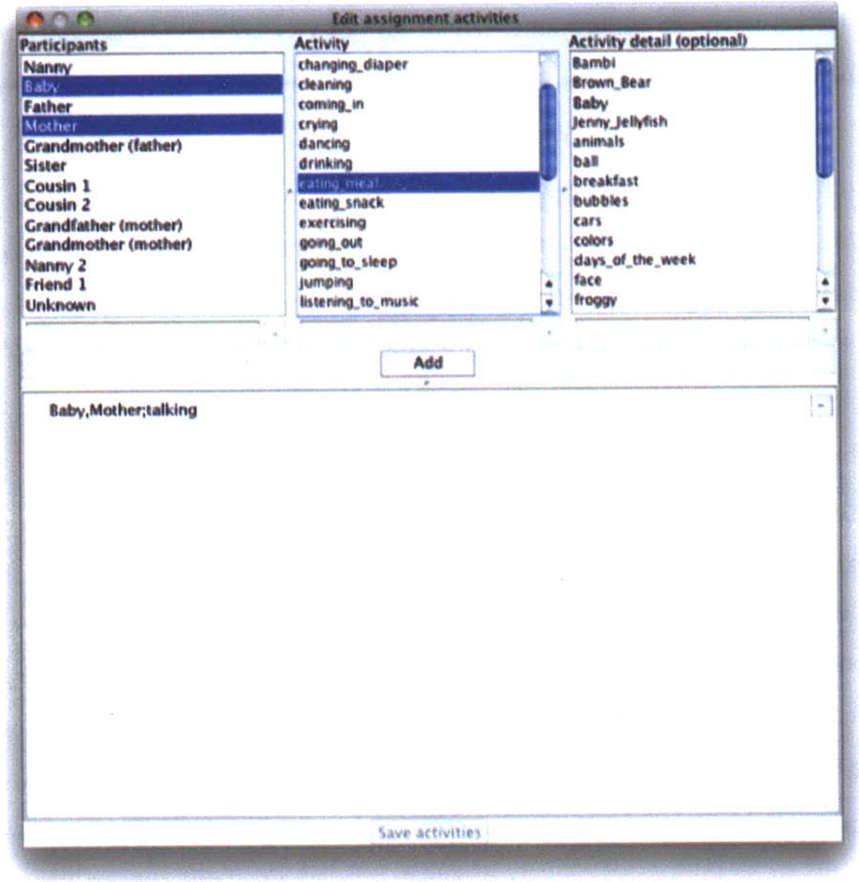
Bruner emphasized studying language acquisition in the "clutter" of life at home, since this clutter provides the raw material for the structured formats critical to language learning. However, Bruner's choice of words also highlights the challenge in studying them: how do we identify and find the structured activities in the clutter of everyday life? The canonical example of peek-a-boo is unfortunately the simplest case, where most episodes of the activity can be identified by a simple keyword search in the transcripts. But for other activities the solution requires a different approach.

6.2.1 Annotation tool and coding scheme

To begin, we incorporated manual activity annotation as part of the transcription process. Human annotators, in the course of transcribing an assignment, have the most direct contact with the data and are in the unique position to determine what is taking place in an assignment. BlitzScribe was extended with a new user interface for activity annotation, with the goal of steering annotators toward consistent activity annotation at the appropriate level of detail while being both simple and flexible. The tool supports associating multiple activities with an assignment, where an activity consists of an *activity type*, a list of *participants*, and optionally, an *activity detail*, which could help further refine the activity type. For example, an annotator might indicate the nanny was changing the child's diaper as `nanny,child;changing_diaper` or that the mother was feeding the child grapes for a snack as `mother,child;mealtime-grapes`, where `grapes` is the activity detail for the `mealtime` activity type. Annotators could choose from an existing list of participants, activity types and activity details, but could also add new entries which would be stored to the central HSP database and become available to other annotators. This helped maintain consistency but also flexibility: if a truly unique, new activity occurred, the annotator could meaningfully encode it. Figure 6-1 shows the user interface, which annotators could pop up while transcribing.

As with any coding scheme, there are difficult representational choices to make, since not all information can (or should) be preserved in the annotation process. In the above examples, it is clear that the nanny is changing the child's diaper, and the mother is feeding the child grapes, yet in our scheme we refer to people as "participants". This is largely because not all activities have an obvious agent-patient relationship. Clearly, some activities vary depending on the participants involved, but by taking a non-subjective stance on activity we hoped that a smaller, core set of activity types would be adequate.

This scheme evolved out of an initial, free-form process of "tagging" assignments. We began by asking annotators to tag assignments by adding a few keywords to the assignment notes field about the activity and who was involved, but with little other guidance. The list of tags differed between annotators, but there was a great deal of overlap after normalizing different names for the same activity. These tags formed the initial basis set of activity types and details for the activity annotation tool. Periodically, through group meetings with the annotators, some activity types were merged or renamed. For example, `eating-meal` and



Baby,Mother;talking

Figure 6-1: The activity annotation tool, an addition to BlitzScribe which allows transcribers to list the activities that occurred in an assignment. To specify an activity, annotators select the participants, a main activity type, and can optionally choose an activity detail. These are first selected from the lists at the top, and pressing the "Add" button associates the activity with the assignment. If necessary, annotators can add new entries to the list of participants, activity types and activity details, which will then be shared with other users of the tool. In this screenshot, the assignment already has one activity and the annotator is in the process of adding `Baby,Mother;eating_meal`.

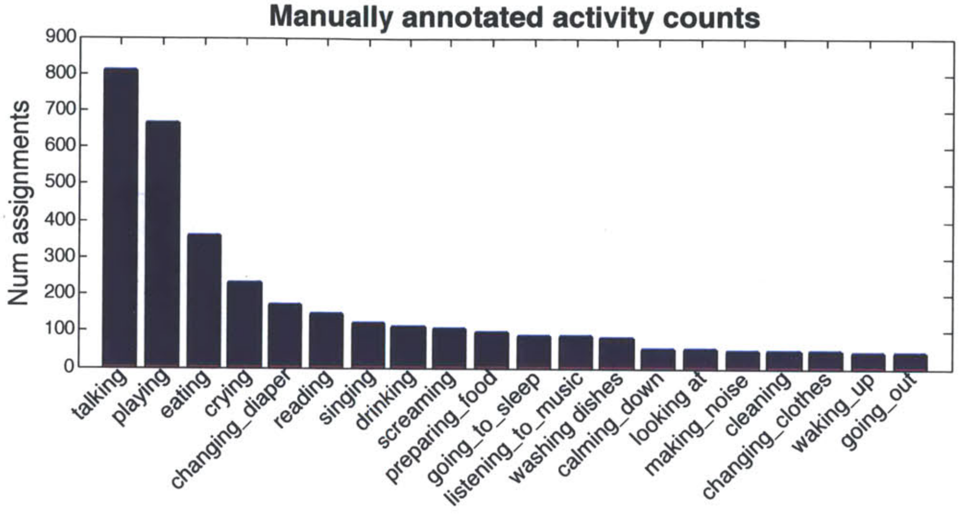


Figure 6-2: Number of assignments labeled with each of the top 20 activity types.

eating-snack were merged since the distinction was unclear. In some cases, activity types were split if an activity type and activity detail combination was consistent and indicative of its own activity. This coding scheme owes much to Carolyn Hsu and Katie Sheridan, who were instrumental in thinking through this process.

6.2.2 Crossmodal characteristics of activities

Using the activity annotation tool, transcribers labeled more than 1400 assignments with roughly 3700 activity annotations. Of the 58 activity types, the top 20 most frequently used activity types were associated with at least 40 assignments. Figure 6-2 lists these activity types and shows the number of labeled assignments. For the most part the activity types are self-explanatory, but the semantics and conventions for use were discussed and documented by the annotators.

It is worth noting that in BlitzScribe, annotators do not have direct access to the location of the activity, nor do they see the video. Their activity annotations are based on their interpretation of *what was happening* in the assignment – a relatively high-level, human description of what transpired during a 15-minute period of time. How do activities manifest across modalities?

We consider three nonlinguistic modalities for activities: time, space, and participants. In addition, we also consider the linguistic modality by characterizing the words associated with activities. Since the granularity of an activity label is at the level of an assignment, then the entire assignment time period is associated with the activity. We are particularly interested in routine, daily activities, so just the *hour* of the assignment is used. This effectively projects activity onto a time range of 0-23 hours (ie. midnight to 11pm). Spatial information is obtained using a similar projection, by taking the primary audio channel for the utterances contained in an assignment and using that to label the attached activities. Of course, utterances may occur in multiple channels, but for simplicity only the one with the most speech is used. Alternatively, the video based methods from chapter 5 could be used for better location estimation. To link activities to participants, each activity annotation explicitly lists the participants involved, or this can be inferred by selecting those speakers that have a significant presence in the assignment. Finally, to explore the relationship between language and activities, we look at the words present in assignments that have a given activity label.

While any activity can be "projected" onto these four modalities, is there a significant link between an activity and a modality? To investigate this, we consider if there is any difference between those assignments that have a particular activity label and those that do not. For example, are assignments with the `changing_diaper` label spatially or temporally different from other assignments? Are certain participants more or less represented in assignments with particular activity labels? To test for a temporal relationship with an activity, a χ^2 test for independence was performed on the number of assignments by hour with and without an activity label. The same approach was used to test for a spatial relationship, using assignment counts by channel id. In the case of participants, a χ^2 test was performed for *each* participant (since multiple participants could be associated with a single assignment.) We do not perform a significance test in examining the linguistic relationship, but score each word according to its likelihood in the activity relative to the other activities and report the 15 top scoring words.

Figure 6-3 illustrates this experiment for three activities: `reading`, `going_to_sleep`, and preparing_food. The χ^2 significance tests are performed with $\alpha = .05$. To obtain the score for word w in activity k, the activity's word distribution β_k is first calculated by counting all occurrences of w in assignments with activity label k and normalizing. Note that for assignments with multiple activities, the count for each word is evenly allocated across the activities (and fractional counts may result). β_k is smoothed to ensure a positive probability for those words with zero counts, since not every word will be observed for all K activity types. The score $s_{k,w}$ for word w in activity k is calculated as

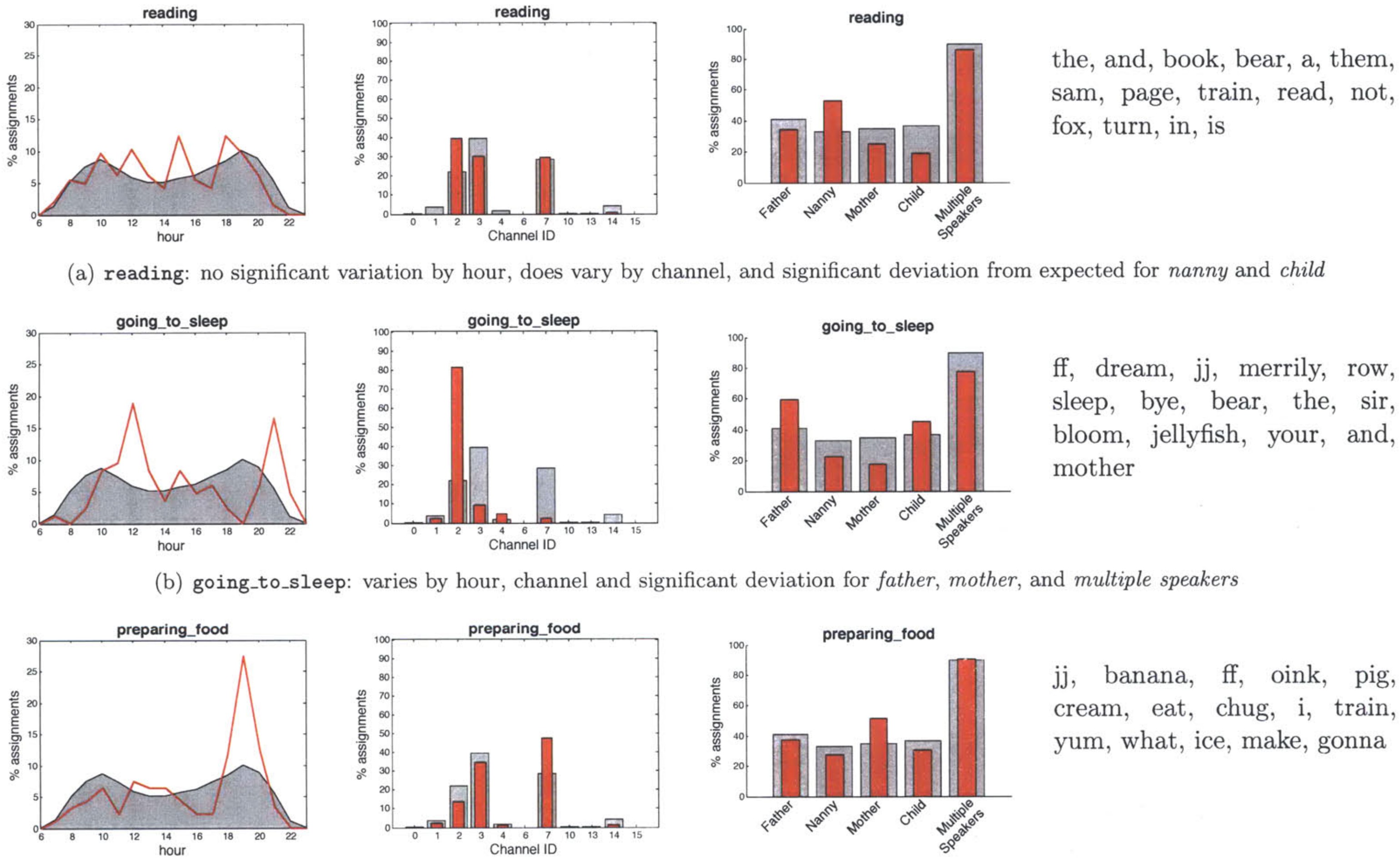
$$s_{k,w} = \beta_{k,w} \log \left(\frac{\beta_{k,w}}{\left(\prod_{j=1}^{K} \beta_{j,w} \right)^{1/K}} \right) \tag{6.1}$$

This score comes directly from (Blei and Lafferty, 2009) and captures both the frequency of the word but also its informativeness relative to the rest of the activities. [1]

6.2.3 Discussion

As figure 6-3 shows, the three activities differ in how they manifest across modalities. The reading activity does not show specific temporal structure but is spatially tied to the child's bedroom (audio channel 2) and shows significant participant variation for *nanny* and *child.* Here, participants were characterized by their utterance counts, and notably the nanny is significantly overrepresented while the child is significantly underrepresented. If the nanny usually reads stories to an attentive (and quiet) child, then this pattern would emerge. The top words include "book", "turn", "page", "fox", "bear", "sam", all of which may relate to the act of reading and to specific stories. The going_to_sleep activity *is* temporally tied, with an expected peak at about 9pm but also a broader peak around noon, perhaps reflecting an after lunch nap. Spatially, this activity is grounded in the child's bedroom and shows above average vocal activity for *father* and *child,* which may indicate that the father usually would prepare the child for bed (anecdotally, this seemed to be the case for the evening

[1] Rewriting this expression as $s_{k,w} = \beta_{k,w} \left(\log \beta_{k,w} - \frac{1}{K} \sum_{j=1}^{K} \log \beta_{j,k} \right)$ shows that if the word has an above average log probability in activity k (the term in parentheses) it will have a positive score, with a higher score if the word has a high probability.



(a) `reading`: no significant variation by hour, does vary by channel, and significant deviation from expected for *nanny* and *child*

(b) `going_to_sleep`: varies by hour, channel and significant deviation for *father*, *mother*, and *multiple speakers*

(c) `preparing_food`: varies by hour, channel, and only significant deviation for utterances by *mother*

Figure 6-3: Relationship between three different manually annotated activities and four modalities: time of day, location, speaker utterance count, and words. For each activity, the distribution for assignments with the activity is compared to the overall assignment distribution. For example, for time of day, the overall distribution of assignments by hour is shown shaded in gray, and the distribution by hour for assignments with the activity is shown in red.

going_to_sleep activities, although probably not the case for afternoon naps). Here, the words include "sleep" and "bye" but also words such as "bear", "dream", and "row". These likely result from bedtime stories and nursery rhymes. Finally, preparing_food is strongly temporally tied to about 7pm, spatially tied to the kitchen (audio channel 7) and the only notable participant is the *mother*. Here, the words certainly include food and mealtime words such as "eat", "yum", "banana", "ice", "cream" but also words like "train", "chug", "oink", and "pig".[2] These last four words may not so much derive from the preparing_food activity as from other correlated activities. In fact, more than half of all assignments labeled with preparing_food were also labeled with playing, which could account for these words.

This highlights an important point that will be further discussed below, and that is the question of how multiple activities – even simultaneous activities – will manifest across the four modalities. With speech in particular, if some portion of the conversation relates to one activity and some to another, then dividing all words evenly across activities may not be the best approach. Instead, if each activity is known to account for a fraction of the assignment, then words could be allocated according to this fraction. Better still, if a language model for each activity is available, then each word could be allocated according to its relative proportion between models. We reserve this idea for the next section on fully automatic methods.

The power of this analysis is not the discovery that mealtimeltime occurs in the evening in the kitchen, or that the child goes to sleep at night in his bedroom. Rather, what is important is that a data driven approach can reveal the high-level patterns of daily life, some of which are expected but some which may be unexpected as well. But does characterizing the patterns of daily life provide any insight into early word learning?

6.3 Linking manually annotated activities to word learning

Although only a fraction of the transcribed assignments have activity annotations – about 10% – we can perform a preliminary analysis linking word usage across activities to the

[2] Although "oink" and "pig" could conceivably relate to mealtime.

child's word learning. The basic idea is that words that are contextually restricted and strongly tied to a small number of activities may be easier to learn. Although formats are highly structured and often "script-like", we suggest that some of the key elements of formats may be subsumed by a more general notion of "predictability" which provides a broad, flexible continuum for connecting language to social activity. Given an activity, what words does one *expect* to hear – how predictable is a word within an activity? And how distributed is a word across different activities? That is, what is the range of activity contexts encompassing a word's use? We focus on characterizing this second aspect.

6.3.1 Characterizing a word's activity distribution

To begin, consider the set of N assignments that have activity annotations ($\approx$ 1400) and the W unique word types across these assignments. Let C_{ij} be the count of word j in assignment i, thus C is an $N \times W$ dimensional matrix. Additionally, assume that the rows are ordered by the timestamp t_i of the corresponding assignment. Next, for the same N assignments, and in the same order, construct an N by K activity matrix A where $A_{ij} = c$ if assignment i has activity label j, otherwise $A_{ij} = 0$. This matrix is normalized so that each row sums to 1 – if there are two activities associated with assignment i, then these two activities will each have a weight of $c = 1/2$.

We wish to consider how caregiver uses of word w are distributed across activities *prior* to the word birth date. Let X_w be a K dimensional row vector with the allocation of counts for word w to activities. To obtain X_w, we consider the subset of rows of C and A that correspond to assignments that are earlier than birth(w), and distribute the count of w in each assignment across the applicable activities. That is,

$$X_{wj} = \sum_{i=1}^{N} \mathbb{1}_{[t_i \leq \text{birth}(w)]} \cdot C_{iw} \cdot A_{ij}$$

where $\mathbb{1}_{[t_i \leq \text{birth}(w)]}$ is the indicator function with value 1 if the assignment timestamp $t_i \leq$ birth(w) and 0 otherwise.

For each of the V words learned by the child, X_w is calculated and collected into the $V \times K$ matrix X, where X_{wj} is just the fractional number of caregiver uses of word w during activity j in assignments *prior to* the birth of w. To give more of an intuition for X, note that summing across the rows of X results in the total caregiver usage count for each word w prior to the word birth. Thus, X is a kind of projection or "splitting" of caregiver uses of the word across activities.

If each row of X is normalized, the rows can be interpreted as the conditional probability distribution over activities for word w in caregiver speech prior to the birth of w, which we denote as p_w. For example, the top three activities in the activity distribution for the word "diaper" are, in descending order, `changing_diaper`, `talking`, and `playing`. This highlights an important factor that must also be considered: the overall probability of an activity. As shown in figure 6-2, `talking` and `playing` are the most frequent activity labels. So the question is, does the activity distribution of the word "diaper" occur across activities in a way that is distinct from the overall distribution of language across activities?

6.3.2 Quantifying the connection between words and activities

To answer this question, we construct an overall activity distribution to serve as a reference point, and then calculate the KL-divergence (relative entropy) for each word's conditional activity distribution against this "background" distribution. If the distributions are identical, then we would not be inclined to argue that the word has any interesting relationship to activity and the KL-divergence will be 0. On the other hand, if the conditional activity distribution for a word varies greatly from the background, the KL-divergence will be large and it could be taken as evidence for an association between the word's use and activity. One caveat is that the KL-divergence of a sampled distribution is dependent on the number of samples, as discussed in appendix A, so this must be taken into consideration.

The overall activity distribution could be constructed by counting and normalizing the use of each activity label across assignments. However, since X is really linking words with activities (rather than assignments with activities), we distribute each assignment's word count across the activities for the assignment. Let $n_i = \sum_j C_{ij}$, the total word count for assignment i, and let $\gamma_{ij} = n_i \cdot A_{ij}$. γ_i captures the number of word occurrences "contributed" to the assignment by each activity. However, it may be preferable to smooth A so that there are no zeros in the matrix and every activity has at least some (small) contribution, so with a small smoothing parameter $\alpha \geq 0$ we have

$$\gamma_{ij} = n_i \frac{A_{ij} + \alpha}{1 + \alpha K}$$

Averaging over γ_i and normalizing yields an overall K dimensional activity distribution γ^*.

One way to think of this model is that each activity contributes a portion of the words to an assignment. If an assignment has only one label, then it is "responsible" for all the words in the assignment, and if there are multiple labels, then in this simple model each activity contributes an equal number of words to the assignment. Of course, a more sophisticated model might allocate words to activities differently, but we reserve that for the next section.

Note that this methodology is nearly identical to the spatial word distribution analysis in section 5.4. In that analysis, a location distribution of caregiver uses of a word prior to AoA was calculated and compared to the overall location distribution of language use by calculating the KL-divergence. The sensitivity of KL-divergence to sample count was first introduced and addressed in that analysis. One way to think of that analysis was as a projection operation: for a particular word, the pre-word-birth caregiver use of the word was "projected" into the spatial domain and compared to the projection of overall word use. In this analysis, words are instead projected onto "activity-space" – somewhat more abstract, but measurable and interpretable.

6.3.3 Relating caregiver word use across activities to word learning

The computations described above result in p_w, the activity distribution for each word, and γ^*, the overall activity distribution. However, since the number of assignments with manually annotated activity labels is only about 10% of all assignments transcribed, some words did not occur a sufficient number of times and were not included in our analysis.

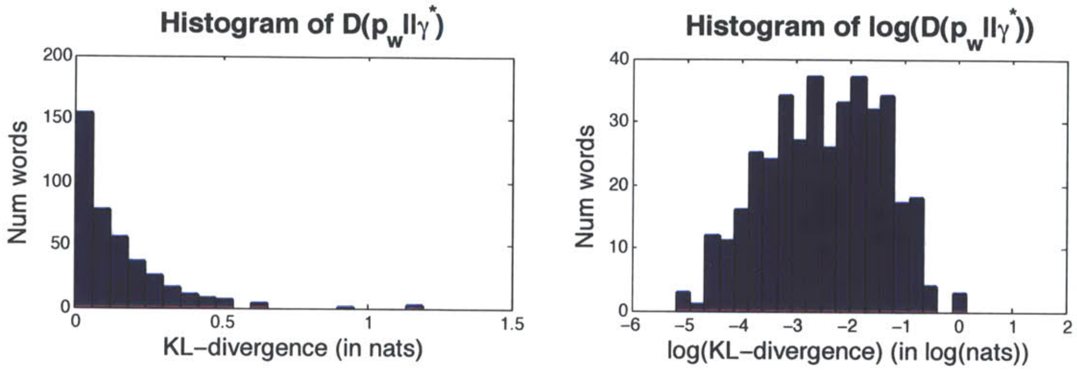


Figure 6-4: Histogram of the KL-divergence between a word's conditional activity distribution and the overall activity distribution, for 394 words in the child's vocabulary that had at least 50 occurrences in the activity-annotated assignments. The KL-divergence is highly skewed; the log-transform de-skews the data.

Filtering out words occurring fewer than 50 times (in pre-word-birth caregiver speech) left 394 of the 669 possible word births for analysis.

For each of these words $D(p_w||\gamma^*)$, the KL-divergence measure, was calculated to capture the similarity between the word's activity distribution and the overall activity distribution. A histogram of these values is shown in figure 6-4. Since this distribution is quite skewed, the log-transformed histogram is also calculated.

Figure 6-5 shows a scatterplot of AoA against the activity distribution KL-divergence for each word. The downward sloping regression line shows that words with more "unique" activity distributions tend to be learned earlier, with a correlation coefficient of $r = 0.30$. That is, words that are used across activities in a distinct fashion, relative to overall language use, tend to be learned earlier.

The KL-divergence of a word's conditional activity distribution also relates to the number of activities in which a word is used. For word i and threshold $t = .6$, the top $k_i(t)$ activities that cumulatively constitute more than $t = .6$ of the word's conditional activity distribution is a crude way of "counting" the number of unique activities in which a word is used. If a word is spread across many activities, $k_i(t)$ will be larger, while if it is strongly linked to only a few activities then $k_i(t)$ will be smaller. Here, we find that words with higher KL-divergence relative to the background tend to occur in fewer activities and have smaller

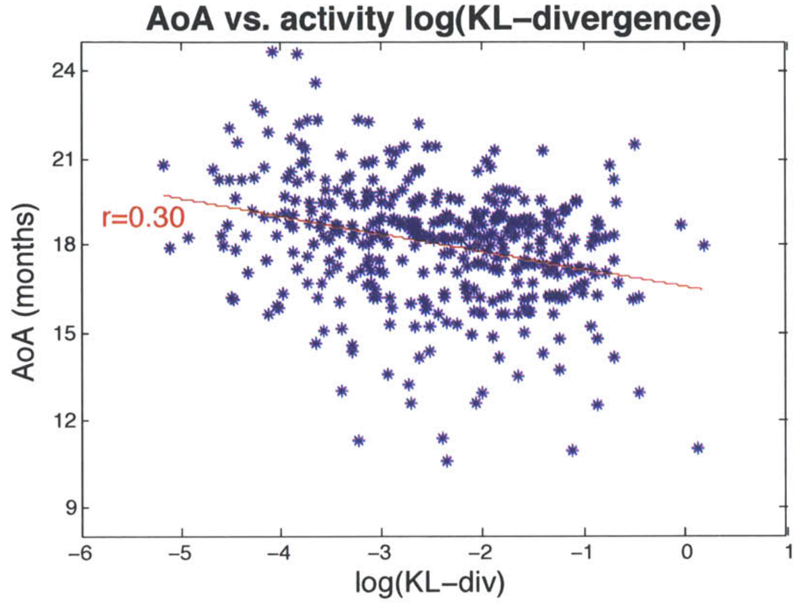


Figure 6-5: Relationship between the log-transformed activity KL-divergence and AoA, for each word. This measures a word's uniqueness in terms of how it is distributed over activities. The negative slope indicates that words that have more distinct activity distributions, relative to the average, are learned earlier.

k_i values, as shown in figure 6-6.

Just as with spatial context, there is a relationship between the number of samples used in constructing the word conditional distribution and its KL-divergence value. Therefore, we calculate the residual log KL-divergence after factoring out the sample count, obtaining the scatter plot in figure 6-7 and a better overall prediction of AoA. Importantly, the same general trend holds: words that are more unique in how they are used across activities tend to be learned earlier, and words that are more unique by the KL-divergence measure tend to occur in fewer activities as measured by the $k_i(t)$ counting method described above.

6.3.4 Discussion

These findings show that how a word in caregiver speech is distributed across activities relates to its age of acquisition. Words that are more "unique" in their activity distribution, which tend to be focused in fewer activities, are learned earlier. This is consonant with the

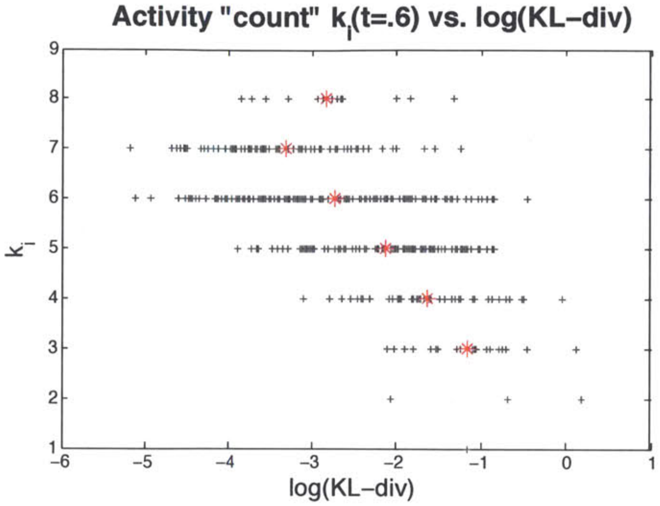


Figure 6-6: Scatter plot showing the relationship between a word's log KL-divergence and the effective number of activities in which the word is used, as counted by $k(t = .6)$. The mean log KL-divergence value for words with each k value is shown in bold red. This plot reveals that words with a higher KL-divergence relative to the overall activity distribution tend to have most of their use concentrated in fewer activities; such words are more contextually specific.

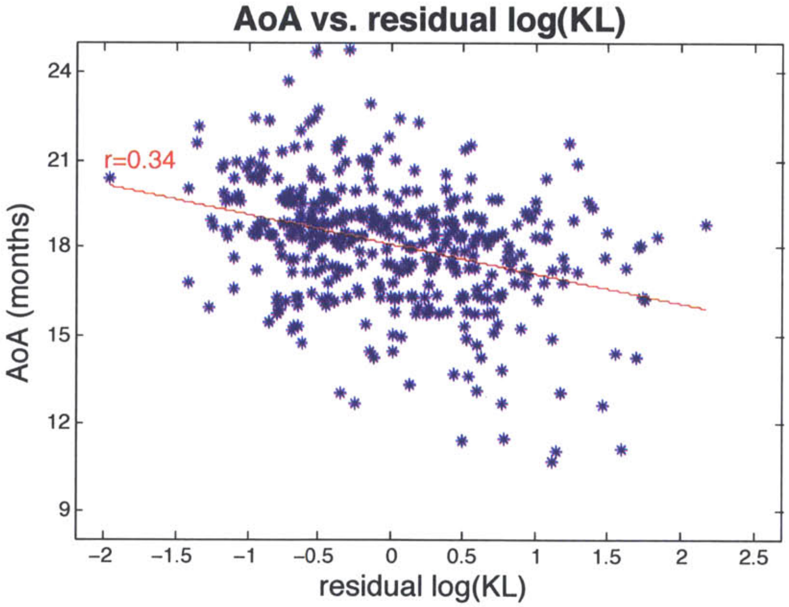


Figure 6-7: Relationship between the log-transformed activity KL-divergence *residuals* (as predicted by log sample count) and AoA. Using the residuals effectively corrects for the dependence of KL-divergence on sample count.

general argument for the role of structured, predictable context as supporting word learning. The finding is also in agreement with the spatial context results presented in section 5.4, and uses a very similar analysis approach.

How do these measures of a word's activity distribution relate to the child's experience? One way they relate is by capturing some aspect of the strength of association between a word and an activity. Although the word "run" is more frequent than the word "kick", "kick" has a more unique activity distribution (even when controlling for frequency) and "kick" is learned earlier. One might imagine the word "kick" is used primarily only in particular activities, such as games, while the word "run" could easily occur across many activity contexts. So from the child's perspective, "kick" may be more salient or its restricted usage may provide a useful constraint.

This analysis showed that the way words are used across daily activities is predictive of word learning. Activities are "high-level" structures, and we relied on human interpretation of the data to provide the labels. Beginning with manually annotated activity contexts – labels for *what is happening* – we found that they project meaningfully onto space, time, language and participants. In the next section, we ask whether we can go in the other direction: can activities be inferred from observed data?

6.4 Automatic methods for identifying activity contexts

The data collected for the Human Speechome Project consists of hundreds of terabytes of audio and video, with a single day typically spanning several hundred gigabytes of data. Yet asking one person to describe the day's events to another might require only a few pages of text. At this level, the description is at once a very lossy form of compression as well as a distilled narrative. The manual activity annotations constitute a kind of structured narrative designed to relate, with some level of objectivity, "who did what" over the course of the day. Relying on such human interpretation of the data is a good starting point for capturing the relevant activities that transpired, but labor-intensive manual annotation

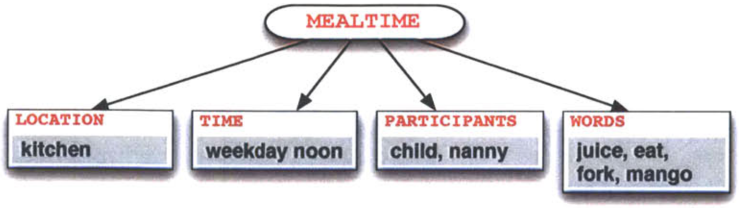


Figure 6-8: Latent variable decomposition for the `mealtime` activity

can also be a limiting factor. Could automatic methods be used to discover the important underlying activities that give rise to the observed data?

Viewing the temporally aligned stream of transcripts, speaker annotations and spatial distributions as jointly explained by an underlying activity suggests modeling activity as a *latent variable*. Figure 6-8 shows an example of how the `mealtime` activity might manifest in terms of *who* is involved, *when* and *where* it takes place, and *what* is said during the activity. In the previous section, correlates across these four modalities were derived from the data for several activities.

6.4.1 Latent variable models

Probabilistic methods for inferring latent variables from observed data exist for a range of applications and data types. For example, Hidden Markov Models are often used to model observed sequential data as a sequence of hidden "state" transitions. Such models have been successful in speech recognition applications, where the hidden states might reasonably correspond to phonemes or syllables, but what is observed is the acoustic data (Rabiner, 1989). Another task is to find the latent topics in a large collection of documents. Articles on a news website might refer to themes such as "politics", "sports", "arts" and so on, and a document that is about one of these topics would contain words characteristic of the topic. This suggests a generative process view: the underlying states could be "phonemes" that generated the audio signal, the underlying topics are word distributions contributing words to the document. A probabilistic representation of documents in terms of latent factors or topics was taken in (Hofmann, 1999), while Blei et al. (2003) proposes a full generative model called Latent Dirichlet Allocation.

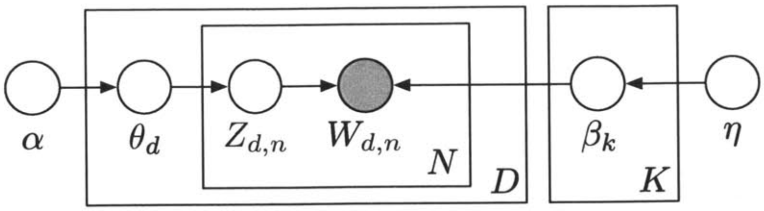


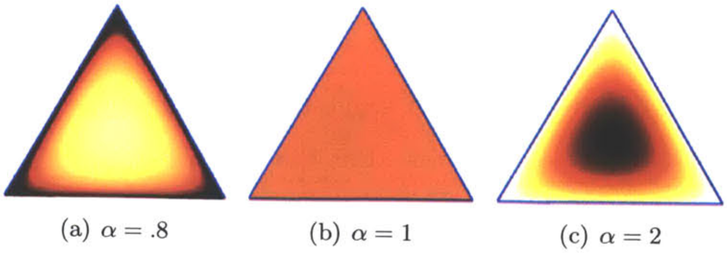
Figure 6-9: The structure of LDA in plate notation. The smaller interior box contains the variables that replicate for words within a document, the outer box contains the variables that replicate for each document in the corpus. Arrows indicate dependencies. For example, the identity of word $W_{d,n}$ depends on the word's topic label $Z_{d,n}$ and the word distribution for the topic.

6.4.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic generative model for discrete data, often used in modeling collections of documents. In the case of document collections, a document is a mixture of *topics*, and topics are in turn distributions over words.

Viewed as a generative model, a document is generated as by first choosing a topic distribution, sampling a topic from this distribution for each word and drawing the word from the distribution for the topic. A document is represented as an unordered collection of words and their counts, a so-called "bag of words" model. The plate notation picture for LDA is shown in figure 6-9, which summarizes the variables in the model and captures the dependency structure.

Briefly, the variables in LDA are structured as follows. For a particular document d with N words, $w_{d,n}$ in the innermost plate represents the observed word at position n. This word depends on the topic sampled for this position in the document, $Z_{d,n}$, and β, the conditional word distributions for the K topics. Thus, β is in a plate that replicates K times with $\beta_{i,j} = \Pr(w_j|z_i)$. $Z_{d,n}$ can be thought of as an index into β for word n in document d. Since $Z_{d,n}$ and $W_{d,n}$ are sampled for each word in the document, they are part of the inner plate that is repeated N times. $Z_{d,n}$ is sampled from a document level parameter θ_d, which describes the distribution over topics for the document. This parameter is part of the outer plate which replicates once for each of the D documents in the collection. Finally, α is a corpus-level parameter for the symmetric Dirichlet prior used in sampling a topic



(a) $\alpha = .8$ (b) $\alpha = 1$ (c) $\alpha = 2$

Figure 6-10: Symmetric three dimensional Dirichlet distribution for three choices of α. Darker colors indicate higher probability, with $\alpha = 1$ yielding a uniform probability distribution.

mixture for each document, and η is a corpus level parameter for the symmetric Dirichlet prior over the word distribution for a topic. The parameters α and η are greater than zero.

Each document d is a distinct mixture over all K topics, with the topic distribution given by θ_d. A "sparse" topic distribution is one in which the probability mass concentrates on only a few topics, which is often preferred for modeling document collections. The parameter α controls the Dirichlet prior on document-level topic distributions, with small α leading to sparser distributions. To give an example of how α affects the Dirichlet prior, which in turn affects the choice of θ, consider an LDA model with $K = 3$ topics. Although θ is a probability vector with 3 components, the components must be nonnegative and sum to 1, thus θ is a point on a two-dimensional simplex that can be represented as a triangle. For a document d that is an even mixture of all three topics, θ_d will be in the center of the triangle, while if the document is exclusively drawn from one topic, θ_d will be at the corresponding corner of the triangle. The Dirichlet prior governs how θ is sampled; $\alpha < 1$ pushes the modes of the Dirichlet prior to the corners, thus leading to sparser topic mixtures. A uniform distribution over the simplex is achieved by $\alpha = 1$, while $\alpha > 1$ moves the mode to the center of the simplex. A three dimensional symmetric Dirichlet distribution for three choices of α is shown in figure 6-10.

Inference in LDA

The generative model perspective of LDA suggests how a collection of documents might be generated by an underlying set of topics. But the real utility of LDA is in inferring the hidden variables from the observed data. Given an LDA model α and β and the observed words $\vec{w}$ in a document, what is the posterior distribution on θ and $\vec{z}$, the allocation of words to topics for the document? The posterior $\Pr(\theta, \vec{z}|\vec{w}, \alpha, \beta)$ is intractable to compute directly, but approximation techniques can be used such as the variational technique presented in (Blei et al., 2003).

A second issue is learning the model itself – in particular, what is β, the word distribution for each topic? In addition, α and η can also be estimated. Blei et al. (2003) derives iterative expectation-maximization procedures for estimating β, α and η. An important parameter that must be supplied to the algorithm is K, the number of topics to estimate for the data. This is not unlike many clustering algorithms, such as k-means, that seek the hidden cluster structure in the data provided. More sophisticated non-parametric techniques such as Hierarchical Dirichlet Processes (Teh et al., 2006) can be used to infer the number of topics.

6.4.3 Using LDA

Our description of Latent Dirichlet Allocation is informal and leaves out many details, but for the present we only wish to summarize how LDA can be used for finding the hidden structure in a collection of documents.

The representation of a document in LDA is as an unordered set of words and counts, a "bag of words". The set of all unique word types in the document collection constitutes a vocabulary with V words, so each document can be treated as a V dimensional vector with nonzero counts for words that occur in the document. For N documents, the entire corpus can be represented as a (sparse) $N \times V$ dimensional matrix. In our experiments, this is the data file format used.

As we have already seen, there are many rare words that may only occur a few times even in a very large corpus, so it is usually best to preprocess documents before applying LDA inference algorithms. One step is to *stem* the words, mapping different forms of the same word to a common root. Stemming usually preserves much of a word's meaning while often significantly reducing the vocabulary size, which can help with data sparsity issues for rare words. We use the popular Porter stemmer (Porter, 1980), which does not necessarily yield well formed root words but is effective at mapping morphological variants to a common root string. Even after stemming, there are some stems that may have an insufficient number of samples for robust statistical estimation, so the vocabulary may be further filtered by removing words (or stems) that occur in too few documents or that occur an insufficient number of times.

Code for the algorithms developed in (Blei et al., 2003) are publicly available online (Blei, 2012) and can be directly applied to a corpus that has been represented as a sparse, bag of words matrix. Running the algorithm requires choosing the number of topics K and an initial value for α. The output consists of α and the $K \times V$ matrix β. In addition, a file containing an $N \times K$ matrix γ is produced, where the rows contain each topic's smoothed, fractional word contribution to the total document word count. Normalizing row d in γ effectively yields θ_d. Finally, a file that lists the word types in each document and the source topic id for the word is produced.

Interpreting an LDA model

One way to interpret the output of LDA is to focus on the topic distributions in β. Topics do not have names, they are simply distributions over the V words in the vocabulary. But inspecting the most probable words can be helpful in understanding what the topic is "about". Blei et al. (2003) gives an example of a topic where the top 5 words are "new", "film", "show", "music", and "movie", which he names "Arts". However, a word that is highly probable in every topic is not particularly informative. Alternatively, the top words for inspecting a topic can be selected for their discriminative power, by combining both the probability of the word and its uniqueness to the topic. Blei and Lafferty (2009) suggests a scoring function to capture this, which we have already introduced in our examination of the words associated with human labeled activities as equation 6.1.

Another way to interpret the output of LDA is to focus on θ_d, the topic distribution for each document. In the previous example, does the label "Arts" represent a common theme for those documents that are strongly associated with this topic? Inspecting the documents that have a strong mixture component for a topic can be very helpful.

Latent Dirichlet Allocation is a powerful technique, but still requires analysis, visualization and interpretation to fully benefit from its use. Extensions and related techniques have been proposed to address correlations between topics (Blei and Lafferty, 2006a), topics that change over time (Blei and Lafferty, 2006b), and documents with associated attributes (Mimno and McCallum, 2008). More detail on theory and applications can be found in (Blei et al., 2003; Steyvers and Griffiths, 2007; Blei and Lafferty, 2009; Blei, 2011) and other sources. However, the basic version of LDA turns out to be a good starting point for investigating activity contexts, as described in the following sections.

6.4.4 LDA topics as proxies for activity

The human labeled activity annotations described in section 6.2 reveal structure along multiple dimensions, including the linguistic dimension. Moving in the other direction, using LDA to find linguistic structure may provide a handle on the underlying activity structure, with topics serving as proxies for activity. In fact, the linguistic structure that we found in the manual analysis of activity had several shortcomings relative to LDA. In the manual analysis, words were associated with activities by evenly splitting word counts in an assignment across the associated activity labels. However, evenly splitting the words does not account for the fact that words are differentially associated with activities, and that activities may be unevenly represented in an assignment. Not accounting for these two aspects blur the lines of an activity's true linguistic structure. LDA accounts for both of these aspects with a per-topic (unigram) language model (β) and a per document topic mixture (θ_d).

In the next section, we explore LDA as a method for identifying activity contexts and linking activity to word learning, using the same visualization and analysis methods developed for manually annotated activities.

6.5 Using LDA to identify activity contexts

To use LDA to identify activity contexts and scale up our analysis of how caregiver speech across activities relates to the child's early word learning, there are several issues to address. The first is how to represent the Speechome data for use in LDA. The second is how to interpret the topics that LDA yields – are they capturing activities? Finally, do they provide a useful way to characterize how caregiver input speech relates to the child's word learning?

6.5.1 Preparing the corpus

We begin by first partitioning the Speechome Corpus into "documents". In its typical usage, a document is assumed to exhibit some coherent and usually sparse topic structure. In our case, we wish to use topics as proxies for activity, so documents are constructed to capture linguistic evidence for one or a few activities. Therefore, we temporally order all transcripts and then partition the sequence into documents that each span a fixed time range using a sliding window procedure.

For this work, each document spans 10 minutes of recorded data, which we might also think of as a 10 minute "episode". For the activities we expect to find such as `mealtime`, `changing_diaper`, and so on, 10 minutes seems a reasonable duration. Episodes at this granularity may fully contain an activity, or if the activity spans episodes it is unlikely to span more than a few. Beginning at the 9-month mark we advance a window over the corpus, shifting the window forward by 10 minutes up to the 24 month mark. All transcribed speech in a window was output as a document, skipping empty time windows that didn't contain speech, resulting in 18,751 documents.

Of course, there are other ways of partitioning the transcripts. For example, the transcriber assignments could be used, but tying this analysis to the unit of annotation work introduces unnecessary complexity. However, a more sophisticated method, such as finding long gaps in speech, room transitions, and other natural boundaries in the data may be worth investigating.

As discussed above, some preprocessing of the documents is often helpful when applying LDA. After normalizing the words with a combination of handcrafted word mappings and the Porter stemmer (Porter, 1980), and then removing words that occurred fewer than 6 times or in fewer than 5 documents, the more than 30,000 word types was reduced to a 6,731 word vocabulary.

The resultant bag of words file contains the word usage counts in each episode, but episodes also have other relevant attributes. In particular, the participants, location, and time of day for the episode is of interest. To represent the participants, the word counts attributed to each speaker are used. Location is represented in a similar fashion, by tracking the word counts by room (using the audio channel information.) For both of these modalities, further discretization is possible if desired. Time is represented explicitly by using the start timestamp for the episodes, but as in our analysis of manual activity, time is collapsed to hour of day (ranging from 0-23) to capture recurring daily patterns.

6.5.2 Investigating activity structure

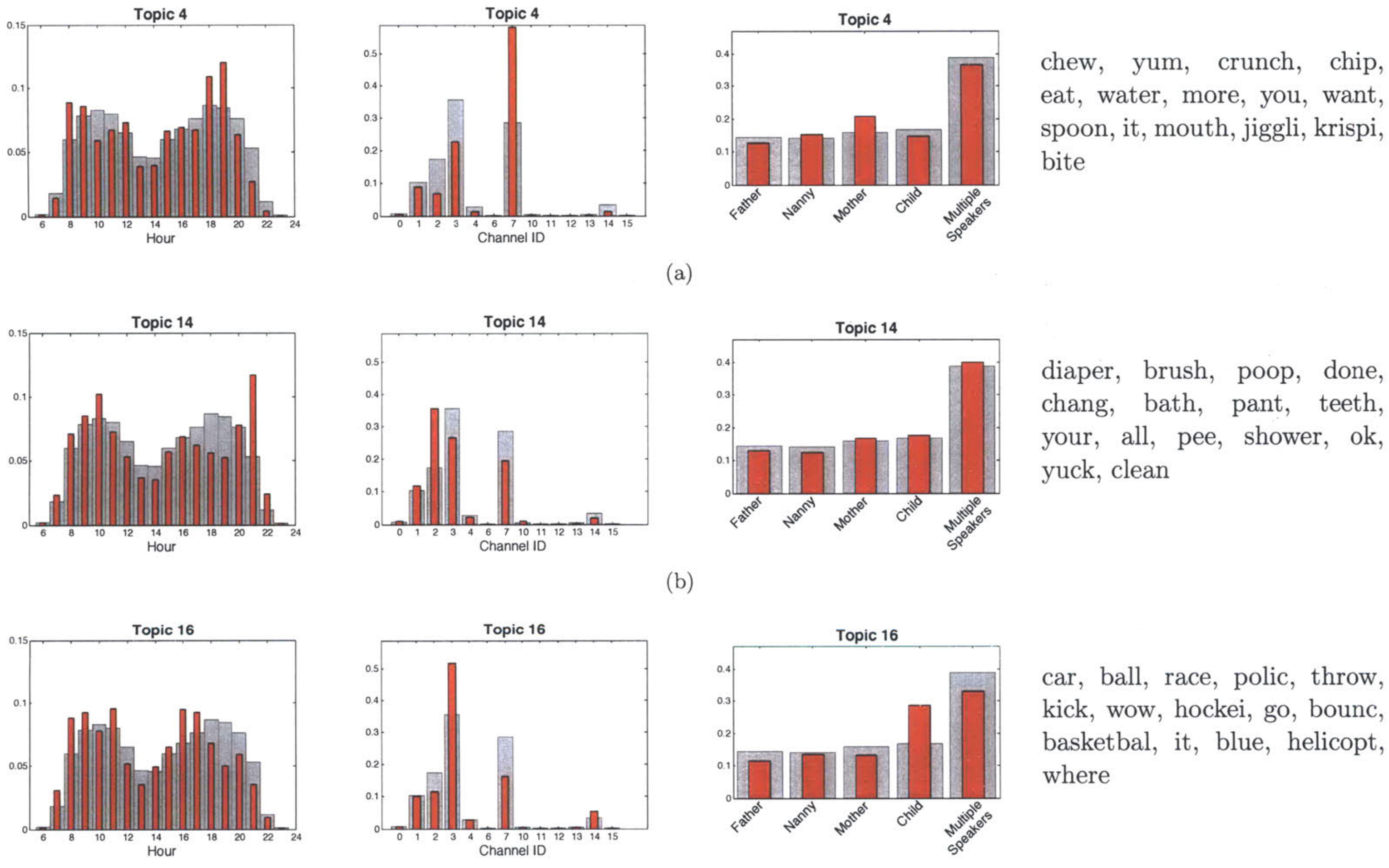
Running LDA requires choosing K, the number of topics to identify. To some extent, this is part of the art in applying this technique, although methods that optimize objective measures are mentioned in (Steyvers and Griffiths, 2007). In this work, we have used the number of human annotated activity types as a guide, and experimented with different choices for K, settling on $K = 25$ in what follows.

LDA also takes an initial value for α, optionally updating this value as the algorithm runs. Starting at $\alpha = 1$, the value decreased to $\alpha = .09$ after 25 iterations. This is a prior that leads to sparse topic mixtures θ_d on the documents.

Figure 6-11 shows a visualization of three LDA topics in terms of their top words as well as their manifestation in the nonlinguistic modalities of space, time and participants. We see a topic that appears to capture mealtime, diaper change (or a more general bathing activity), and a playtime activity. Topic 4, the mealtime topic, peaks in the morning and early evening and in the kitchen, where meals are usually eaten. Topic 14, the diaper change/bathing topic, generally peaks in the morning and late evening and in the baby's bedroom, where the diaper change station is located. However, neither of these topics seem strongly associated with any particular speakers. Topic 16, which looks like a playtime topic, is peaked in the 7-11 AM range and in the mid-late afternoon, is strongly tied to the living room, and the child is a dominant speaker.

Of course, not all topics are directly interpretable as activities, and many topics only capture general patterns of language use. A more direct way to assess whether a topic is capturing a coherent activity (or activities) is to link the topic labels to activities that have been manually annotated. This can be accomplished in a straightforward manner by applying the LDA model to the assignments that have human produced activity annotations, identifying the primary topics for each assignment, and correlating topic ids with activity labels. More specifically, for each of 1464 assignments with activity labels, LDA inference is performed to obtain θ_d, the distribution over topics for the assignment, and then discretized to select the top topics that account for 70% of the total probability mass in θ_d. The pairwise correlation between the topic ids and the activity labels is computed, considering only those correlations that are significant at $\alpha < .05$. While the correlation coefficient may be negative between a topic and an activity label, we are only interested in positive correlations – the cases where the topic's presence indicates the activity label may also be present.

For the three topics examined in figure 6-11, we find that topic 4 is only significantly positively correlated with `eating` and `drinking` (at $\alpha = .05$). If we threshold the correlation coefficient to choose only the activities that are within 80% of the maximum (and also significant), only the `eating` activity remains. For topic 14, `changing_diaper`, `changing_clothes`, `waking_up` and `crying` are the only significantly positively correlated activities, and `changing_diaper` has the highest correlation while the others are removed



(a)

(b)

(c)

Figure 6-11: Time of day, location, participant, and word distributions for three topics that are indicative of mealtime (topic 4, top row), diaper change or bathing (topic 14, middle row), and playtime or games (topic 16, bottom row). Red bars show the topic conditional distribution across time, location, and participants, overlaid on the "background" marginal distribution shown in gray.

by the 80% threshold. For topic 16, the only significantly positively correlated activity is `playing`.

As mentioned above, not all topics capture a coherent activity, and the correspondence between topics and activities need not be one to one. For example, topic 3 is significantly correlated with many activities, but the dominant one is `talking`. This label is both the most frequent as well as the least coherent – much of everyday life is spent lounging at home and chatting with other family members, rather than part of a structured activity. An inspection of this topic reveals that it is peaked in the morning and evening, in the kitchen and living room, with the father and mother as the dominant speakers (along with many "multispeaker" utterances) while the child and nanny are underrepresented. The top 20 words include "i", "she", "it", "that", "like", "know", "just", "year", "think", "thing", "dollar" – all indicators suggestive of adult conversation before and after work.

The `reading` activity is an example of an activity that aligns with multiple topics, rather than having a clear correspondence with a single topic. An inspection of the significant, positively correlated topics reveals that they tend to capture reading, story, and nursery rhyme related terms such as "star", "twinkle", "moon" (topic 13), "fish", "turtle", "cat", "hat" (topic 18), and "bear", "mary", "lamb", and "humpty" (topic 12). Although this illustrates that the structure derived using LDA need not agree with what humans would provide, it is interesting that this set of LDA topics seems to be a refinement of the broader `reading` activity label. This suggests a possible extension to a hierarchical model, in which a topic such as `reading` is defined in terms of subtopics for various books and stories. It also suggests that the `reading` activity has a larger lexical variability than an activity such as `changing_diaper`, which is only significantly positively correlated with topic 14.

6.5.3 Relating automatically identified activity contexts to word learning

The analysis of word learning using manually annotated activities, described in section 6.3, provides the template for the analysis presented here. The outline of the analysis is to calculate the spread of a word in caregiver speech across topics prior to the word birth.

That is, when the child hears a word in caregiver speech, what is the range of activity contexts (as approximated by LDA topics) for these exposures?

The activity context distribution for a word w is obtained by considering the episodes that contain w up to the word birth. As before, C is the $N \times W$ matrix where C_{ij} is the count of word j in episode i. Distributing the occurrences of word w proportionally across topics to obtain X_w is accomplished as

$$X_{wj} = \sum_{i=1}^{N} \mathbb{1}_{[t_i \leq \text{birth}(w)]} \cdot C_{iw} \cdot \theta_{ij}$$

which is then normalized to obtain p_w, the pre-word-birth topic distribution for w.

To quantify the link between a word and activity, we again consider the activity conditional distribution for the word relative to the overall activity distribution. This is to address the fact that some activities (ie. topics) are more strongly represented than others. The overall activity distribution is obtained by considering the contribution, in words, of each activity to each episode. For episode i, consisting of $n_i = \sum_{j=1}^{W} C_{ij}$ words, this is

$$\gamma_{ij} = n_i \cdot \theta_{ij}$$

Averaging over γ_i and normalizing to 1 yields an overall K dimensional activity distribution γ^*.

Once again, the spread for word w over activity contexts is quantified using the KL-divergence between p_w and γ^* as $D(p_w||\gamma^*)$. Words that follow the overall activity distribution have low KL-divergence, while words that are more uniquely tied to activity have a higher KL-divergence. Figure 6-12 shows a histogram of the KL-divergence values, which are highly skewed, and the the more symmetric log transformed KL-divergence values.

The relationship between a word's log-transformed KL-divergence and age of acquisition is presented in figure 6-13. Two plots reveal the same general finding: words with a more distinct distribution across activity contexts, as quantified by KL-divergence, are learned earlier. Since the KL-divergence depends on the number of samples used in calculating

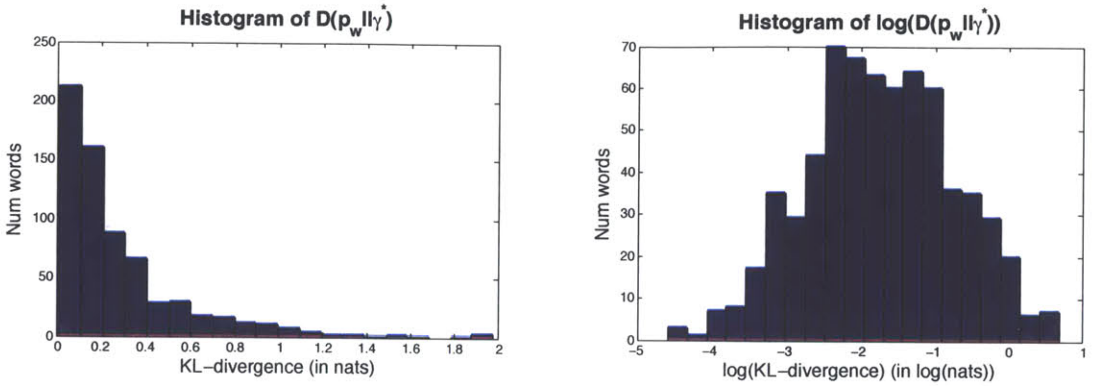


Figure 6-12: Histogram of the KL-divergence between a word's conditional activity distribution and the overall activity distribution. Since the KL-divergence is skewed, the log-transformed version is also calculated which is fairly symmetric.

the measure, as discussed in appendix A, the analysis is also performed using the sample count controlled KL-divergence, revealing a stronger connection between how words are distributed across activity contexts and when the word is learned. Accounting for the relationship between sample count and KL-divergence was also an important part of our analysis of manual activity annotations and a word's spatial distribution. Factoring out sample count better illustrates how a word's use across activity contexts relates to when it is learned.

In earlier work, we investigated the link between a word's topic distribution across the entire corpus and its age of acquisition (Roy et al., 2012). In that analysis, a word's topic distribution was derived from the LDA model directly, rather than the subset of episodes prior to the word birth. The β matrix of conditional word probabilities for each topic was transformed to obtain the conditional topic distribution for each word, assuming a uniform prior on topics. Rather than calculating the KL-divergence relative to the overall topic distribution, the conditional topic entropy was calculated to quantify the extent of a word's spread across activities. That is, $H_w = -\sum_j \Pr(t_j|w) \log \Pr(t_j|w)$ where $\Pr(t|w) = \frac{\beta_{tw}}{\sum_{j=1}^{K} \beta_{jw}}$. The main finding is that, after controlling for frequency, words with lower topic entropy are learned earlier. That is, words that are used across *fewer* activity contexts, as approximated by LDA topics, are learned earlier, consistent with the results presented above.

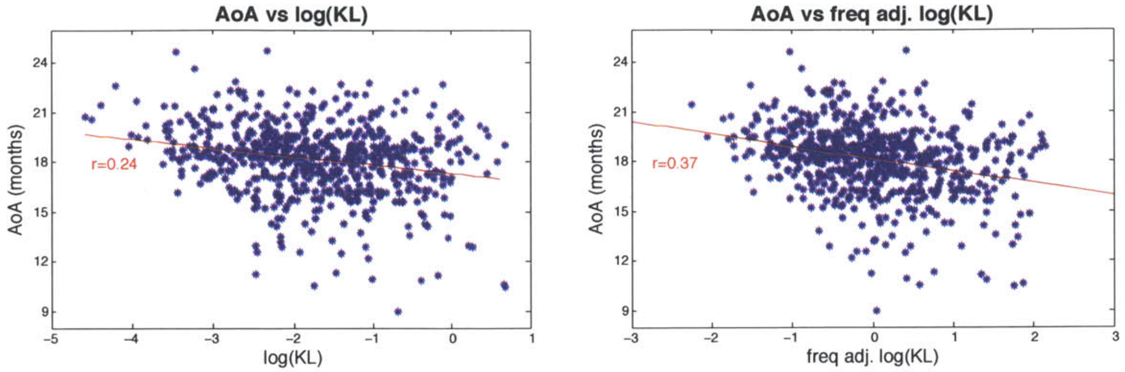


Figure 6-13: Scatter plots and best fit regression lines for AoA on KL-divergence using the automatically identified activity contexts. The plot on the left relates the log-transformed KL-divergence measure for a word's activity context distribution against the AoA, with a correlation coefficient of $r = 0.24$. The negative sloping regression line shows that words that have a more distinct distribution over activities tend to be learned earlier. The plot on the right controls for word frequency, revealing the same effect but a stronger connection between a word's activity context distribution and AoA, with a correlation coefficient of $r = 0.37$.

6.6 Conclusion

This chapter began by arguing that early word learning is dependent on, and facilitated by, the rich context of everyday life at home. Beginning with ideas such as Bruner's *formats*, a characterization of the child's early experience in terms of *activity contexts* was proposed. Activity contexts are labels capturing *what is happening* over time, and they were identified using both manual and automatic methods. Relating a child's exposure to words to the substrate of activity contexts resulted in a strong link between how words are distributed across activity contexts and when they are learned. The key finding is that words that are unique in how they are used across activities, and generally, words that are used across fewer activities tend to be learned earlier.

This result fits into a picture of word learning in which decoding the wide variety of environmental "inputs" is facilitated by contextual constraints. Words that are more context-bound may be more immediately salient to the child, or their meaning may be more tightly linked to other aspects of experience. For an associative learning model, a word that is strongly tied to an activity may primarily co-occur with only a limited range of actions, objects and sensations providing more focused co-occurrence statistics for associative learning (Yu and Smith, 2007; Smith and Yu, 2008). For a more active learning model, such as a model in which the child forms hypotheses about word meanings (Medina et al., 2011; Trueswell et al., 2013), context-bound words have a limited hypothesis space.

Chapter 7

Conclusion

This thesis presented a naturalistic, longitudinal study of one child's early word learning. We have focused on the dynamics of the child's vocabulary growth and the role played by environmental factors. While studies of word learning in a laboratory have contributed significantly to our understanding of learning mechanisms, they have less to say about natural developmental patterns or the environmental structures that can provide footholds into how to use words and the way that language works. To address these issues, our study of word learning is conducted "in the wild". Through analysis of extensive audio and video recordings collected from the home of a young child, we characterized aspects of the child's input as well as his linguistic development to assess their relationship. This descriptive goal is a major thread running through our work.

The basic element of our analysis is the "word birth" – the child's first use of a word in our data. More technically, this is the age of acquisition of a word in his productive vocabulary. But we take the view that words aren't simply "acquired" from the ether, but rather gestate and take root in the rich ground of the child's early experience. The set of word births identifies not only the child's productive vocabulary but also the timeline of its growth. This vocabulary growth timeline was the centerpiece for our analyses – we investigated both the word births themselves and their environmental predictors, assessing the relative importance of environmental factors by their predictive power.

We have approached our study with an emphasis on the role played by the rich context of everyday life at home. One way in which we formalized this idea was by identifying activity contexts to capture the routine activities that make up a child's daily experience at home, and linked the child's exposure to words to these activity contexts. In characterizing the child's early environment, activity contexts are a relatively high-level variable. More directly measurable variables such as the spatial and temporal distributions of word use in caregiver speech, or simply the total exposure to a word also help characterize the child's experience with language. In this thesis, the basic analysis framework has been to operationalize the measurement of variables of interest for words in caregiver speech and to correlate these variables with when the word is learned. We use the predictive power of environmental factors as a means for assessing their relative importance in word learning.

7.1 Contributions

In brief, this thesis has contributed a picture of one child's lexical development. This picture revealed some surprises, in particular, the "shark's fin" shape of vocabulary growth rate in which word learning rate accelerated and then dropped markedly. Models for this growth curve suggested that it is not a sampling artifact, and a preliminary exploration of the child's combinatorial speech suggested that learning new words was giving way to producing new word combinations. In a sense, the child's overall expressivity continued to increase even as his mode of linguistic expression changed. Another surprise was the size of the child's productive vocabulary; with more than 600 words, it is well in the upper ranges of the (Fenson et al., 1994) study. Certainly, from our single sample we must be careful in making generalizations. But it seems reasonable to assume that prior studies, based either on checklists or significantly less recorded data, would be biased toward estimating smaller vocabularies in young children.

Children are not born knowing words; they learn them from their environment. This thesis contributes to our understanding of how the environment shapes word learning. We investigated four variables that characterize the child's exposure to words, beginning with a coarse measure of word exposure (frequency) and a more subtle measure of the temporal usage distribution of a word (recurrence). Here we find that words used more frequently by caregivers are generally learned earlier, but a better predictor is recurrence – words that are used in clusters or bursts tend to be learned earlier. Both of these variables derive directly from the transcripts, but one of our central concerns is how the child's linguistic experience is situated in the broader context of his early life. Therefore, the next variables we considered were a word's spatial distribution and its distribution across activities. In both cases, we found that more contextually constrained words were learned earlier.

This thesis makes methodological contributions to large-scale data annotation and analysis, driven by the challenge of working with the dataset collected for the Human Speechome Project and the kinds of scientific questions that have motivated this work. To fully exploit the scale of the audio and video recordings, to build a detailed picture of the child's vocabulary growth, and to find linguistic and behavioral patterns requires a substantial amount of annotated data.

A key methodological contribution is our approach to speech transcription. BlitzScribe, and the ecosystem of tools for managing a large-scale speech transcription task, may prove useful for other projects that depend on extensive speech transcripts. BlitzScribe is significantly faster than fully manual tools for speech transcription, making previously impractical projects feasible. But scaling up from tens of hours to thousands of hours of audio has presented some significant management challenges. The ecosystem of tools that have evolved around BlitzScribe address these challenges, and are used for distributing work, assessing quality, and tracking progress. BlitzScribe is semi-automatic, but even as fully automatic methods improve, keeping humans "in the loop" will still be important in many projects, especially for research in fields such as cognitive science and developmental psychology. This human-machine collaborative approach made its way into other aspects of our analysis, such as identifying word births using noise-robust statistical methods coupled with efficient human review with a specialized tool.

The direction we have pushed our scientific inquiry has been toward modeling the child's *experience* with language, which required annotations beyond speech transcripts. In particular, modeling contextual variables such as the spatial and activity contexts of word use required both manual and computational methods. In this vein, we contributed activity contexts as a relevant variable in word learning and scalable methods for finding them. Identifying activity contexts was first accomplished with an extension to BlitzScribe for labeling activities according to a standard scheme, which maintained flexibility but supported sharing and reuse of activity labels across multiple annotators. This process added little overhead to transcription, but it was also impractical to label the entire transcribed corpus in this manner. Therefore, a second, fully automatic approach was presented in which activity contexts were treated as latent variables and "discovered" using Latent Dirichlet Allocation. This thesis demonstrated the feasibility of using a fully automatic method, such as LDA, in identifying the activities of daily life and presented approaches to assessing them across multiple modalities.

7.2 Discussion and implications

The descriptive goal of characterizing early word learning and the goal of incorporating the rich context of everyday life into our study are supported by the naturalistic data collected for the Human Speechome Project. In contrast, laboratory studies are often more directly oriented toward the mechanisms of word learning through careful manipulation of variables. For example, the shape, color, and texture of objects, the sentence frame or prosody of speech, or other variables may be manipulated to identify what features are relevant to the child.

In our case, we do not manipulate variables in the child's experience, but identify potentially relevant variables, operationalize their definition and measurement, and relate them to word learning. In a direct sense, we are characterizing the environment, not the learner. But to learn something requires both a learning mechanism – the ability to identify patterns and make inferences from experience – and a "learnable structure" provided by the environment. So by characterizing the learnable structures, and in addition, ranking their importance to the learning task, we can infer something about the underlying learning mechanisms.

7.2.1 Interpreting environmental variables

The four variables – frequency, recurrence, spatial distribution, and activity context distribution – each capture a different aspect of caregiver word use. Each of these variables relates to when words are learned, although their predictive strengths vary. What do we extrapolate from the variables we have measured?

Frequency

Measuring a word's frequency in the child's linguistic experience gets at a basic idea of word learning. To repeat Paul Bloom's comment (2000, p. 90), "People cannot learn words unless they are exposed to them". If the child never hears a word, he won't learn it. But as it happens, hearing a word more frequently correlates with earlier acquisition. This sits well with most learning theories – more exposures provides more opportunity to make the connection between a word and its meaning. More exposures presents more opportunities for speech segmentation processes to isolate the word. Consider the simple associative model of St. Augustine: with limited attentional resources, more exposures to a word means a greater chance of attending to the appropriate object when the word is used. In cross-situational learning, each exposure potentially reduces ambiguity by carving out the common elements of word use across situations, helping to isolate word-referent pairs (Siskind, 1996). Each exposure contributes additional co-occurrence statistics for associative learning (Yu and Smith, 2007; Smith and Yu, 2008) or presents an opportunity to revise or discard a hypothesis about the word or how it should be used (Medina et al., 2011; Trueswell et al., 2013). It should not be surprising that frequency relates to word learning, as it characterizes a basic aspect of the interface between the learner and his environment: the degree of exposure to a word.

Recurrence

But frequency is the weakest predictor in the ensemble of variables we have considered. Instead, in the purely linguistic domain, a word's recurrence better predicts its age of acquisition. Recurrence measures how clustered a word is in time; a high recurrence word is one that, when it is used, is used repeatedly over a short duration. For learners with a limited working memory, a word with high recurrence may occur frequently enough in a short duration to take hold in memory. This may apply at the acoustic level – a repeated sound pattern may be evidence for a lexical unit – but finding repeating patterns is only tractable if the scope is limited, as was the case for the CELL model (Roy and Pentland, 2002). Similarly, if sense-memory degrades with time, then pattern matching may only be effective for sound segments that repeat within a short temporal window. Along the same lines, even if the child's sensory experience (acoustic, visual and otherwise) were neatly pre-segmented into units (words, objects, actions, etc.) the word learning problem would still remain. Even restricting the task to learning object names, the matching problem between words and objects would still be challenging with a perfect, long-term memory – there would be too many candidate words and objects to match.

But in our treatment of recurrence in chapter 5, we suggested another perspective on this variable, although it is not necessarily contrary to the short-term memory view. This idea has to do with recurrence as an indicator of salience in a situation. We gave the example of the use of the word "ball" when a ball is present – it may be used frequently during a period of playtime with a ball, but not used otherwise. Or if the child and caregiver are eating a mango, the relatively rare word "mango" may enjoy frequent use, at least during mealtime. Words that exhibit such "bursty" behavior may be indicative of some interesting underlying cause. Barabási (2010) provides many examples of the "burstiness" of human behavior, with short periods of intense activity surrounded by longer durations of relative quiet, an indicator of non-random behavior. For our case, recurrence may be an indicator that something interesting is taking place in which the word plays a starring role. If the word is relevant to some particular structured activity or recurring episode in the child's experience, then following (Bruner, 1983), the child may be better engaged and the word may be better grounded in other aspects of the child's experience.

Taken together, recurrence may be doing the part of frequency with respect to short-term memory, as well as indicating the salience of a word in some behavior or activity.

Spatial context

Few would dispute that words are grounded in experience and are not used randomly. Rather, words are used in a highly structured fashion. Our investigation into spatial context characterized how a word was used throughout the child's home, by comparing its spatial distribution to the spatial distribution of all words in aggregate. Using KL-divergence to quantify the disparity between the spatial distribution of a word's use and the overall spatial distribution of language in the home, we found that more spatially distinct words are learned earlier. This reproduces the results from (Miller, 2011), and is consistent with the effect found in (Shaw, 2011).

What does a high KL-divergence for a word mean? To be precise, it means that the word is used in a *spatially unique* way relative to the average. One upshot of this is that a word with a high KL-divergence will have certain locations where it is less likely to occur than the average, and other locations where the word is more likely to occur. In contrast, a word with very low KL-divergence doesn't have associated regions of above or below average use. A second aspect of high KL-divergence words is that, practically speaking (though not necessarily), they tend to be used in a more *localized* fashion – they are strongly tied to space.

How might this matter for the child learning words? One speculation is that spatially tied words are more constrained, and more constrained words are easier to learn. Consider a word that refers to an object. If the word is always used in a certain location, then there are likely to be fewer potential referents at that location than if the word were used anywhere in the house. Or consider a particular activity that is tied to a location in the house. A word that is only used at that location is likely to play an important part in that activity, leading back to notions from Bruner. From this vantage point, KL-divergence is measuring a word's scope or "groundedness", with the idea that more grounded words are more strongly tied to other aspects of experience and are more tightly woven into the child's understanding.

But the KL-divergence for a word's spatial distribution can also be considered from another perspective. As mentioned above, a word with a high KL-divergence will have certain locations where it is more likely than average, and other locations where it is less likely than average. Although we measured the conditional location distribution when a word is used, this can be inverted using Bayes' rule to show that the conditional word probability given a location may be above or below the marginal word probability. So it should not be surprising, nor require much mathematical rigor, to believe that the likelihood of hearing a word may increase or decrease depending on the location; the word "blender" is more likely in the kitchen than elsewhere. But this effect of location is more pronounced for some words than others. So with this in mind, perhaps being in a certain location sets up the child's expectations for hearing a word. There may in fact be a small set of words that are expected to occur at this location beyond their normal use, and such expectations may guide and focus the child's attention. Conversely, a word that is spatially distinct will have locations where it is *unexpected*, and one might be surprised to hear it in that context.

Our analysis cannot distinguish between these possibilities, but the notions of constraint, expectation, predictability and surprise are all intimately related and their role in word learning needs more careful investigation.

Activity contexts

Much of the logic outlined above for spatial context carries over to our consideration of activity contexts. We suggested that spatial context is a proxy for activity – that certain activities may only occur in certain locations, and thus a highly spatially-grounded word may really be a word that is strongly tied to an activity. But if a word's link to routine activity is what matters, then we should try to identify and analyze activities directly, which was the subject of chapter 6. We found that words that are used in a more unique manner across activities tend to be learned earlier, analogous to the finding for a word's spatial distribution. As with space, we found that words that have more unique activity distributions also tend to be used in *fewer* activities, although this is not necessarily a requirement of the KL-divergence measure. So what does this relationship suggest about the child's learning mechanisms?

The two overall interpretations for spatial context apply here as well. The first is that a word with a more unique distribution over activities, and which tends to occur in fewer activities, implies a more restricted range of use. The range of situations for a word's use is more limited and narrower in scope, it is generally more *grounded* in a recurring activity. If a word is grounded in an activity, then an understanding of the activity can provide the substrate for an understanding of the word – there is a substantial overlap between the word's use and the activity. But as with spatial context, there is another perspective on the implications of a high KL-divergence across activity contexts. Having a high KL-divergence across activity contexts means a word is more likely to be uttered when some activities are taking place and less likely in others, whereas a word with low KL-divergence isn't much affected by the current activity. Now consider the child engaged in an activity: if any of the subtle statistical sensitivities that children have exhibited in other aspects of language carry over to the present case, the child may have expectations about the words their caregivers will use. Rather than contending with the full range of all possible words, they may be tuned in to the use of a smaller set, a kind of "keyword spotting" approach to the situation. Or high KL-divergence words may be *unexpected* in the current activity, and hearing the word may be notable due to its surprisal. As before, our analysis cannot distinguish between these cases, and they are all related. But it seems likely that a word's complexity in terms of its range of use across activities and the child having some expectations about words given the activity will both contribute to when a word is learned.

Predictability

KL-divergence is actually a coarse measure for the questions we are asking. But the emergent pattern that connects spatial and activity contexts to word births motivates further investigation to tease apart the issues of constraint, complexity and predictability of word use. Recurrence may also be an indicator of a word's predictability – for certain words, hearing them once implies hearing them again soon after. Built into the idea of a word's predictability is a time frame: what is the time period during which we should expect to hear a word? For recurrence, the time frame is specified – it is on the order of a minute. For spatial context, it depends on how long one is in the relevant location. For activity contexts, it depends on how long the activity is taking place. It is only in the case of frequency that the time frame is least tied to immediate experience; the time frame is a day. When the learner wakes up in the morning, he can reasonably expect to hear a word at some point during the day according to its frequency, but this seems like a weak signal if learning is facilitated by prediction. That spatial followed by activity contexts are the best predictors of when words are learned is indicative of the power of context in word learning, whether to set up expectations in the learner as a driver of attention, or as a way of assessing the complexity of a word's part in early experience.

7.3 Future work

The work in this thesis is in many ways a starting point. For the study of language acquisition in the Human Speechome Project, we have prepared the Speechome Corpus which we hope will support a variety of future analyses. This work is an early example of using "Big Data" in Developmental Psychology, and we hope the methods and tools will be further refined and extended. Whatever the case, it seems likely that studies of language acquisition involving terabytes of data will only become more common.

There are many potential next steps; we outline a few possibilities below.

7.3.1 How are words used?

The birth of a word is the focal point of this thesis, with most of our analyses aimed at understanding what contributes to some words being learned before others. In a sense, our focus has been on the "gestation" period for a word, but if we push the analogy further, what is its developmental trajectory after the word birth? Studies such as Dromi's (1987) considered the child's acquisition and use of a word, particularly whether it is used appropriately and if the child has converged on the correct word meaning. Dromi examines the word's *extension*, that is, whether it is applied to only a subset of the appropriate category (under-extension), a superset of a category (over-extension), or if it maps appropriately onto the category (regular extension). For example, if the word "dog" is only applied to a particular toy dog it would be under-extended, while if it is applied to all four-legged creatures it would be over-extended.

In principle, such an analysis is possible with our data through careful interpretation of the audio and video when the child uses a particular word. But this is still a difficult, labor-intensive analysis. However, there are other approaches to investigating the child's use of a word after it is learned. One way this could be explored is by looking at the range of contexts in which the child uses a word. The immediate linguistic context – the words that surround the target word in a given utterance – may shed light on the child's grammatical development. Is the word being used appropriately, and is it fulfilling its generative potential through combination with other words? We touched on this in our discussion of bigrams in chapter 4. We have strived to identify other contextual variables to characterize the child's environment, such as spatial and activity contexts. Just as these were used in assessing the contextual variability in caregiver speech, they could also be used to study the child's speech. Here, however, we may wish to explore the change in contextual use of a word. Of particular interest would be whether a word is only used in particular situations and if this changes with time. One anecdotal example from caregiver speech is the use of words such as "hi" and "bye", which generally occur at the top of the stairs that lead to the front door of the house. The social routines of greeting and saying goodbye have a spatial component – how does the child's use of these words align with that of his caregivers?

7.3.2 Activity contexts

A major target for our work has been to study word learning with respect to the context of everyday life at home, with the view that daily activities structure a child's early experience and may provide the anchor point for a word's meaning and a structure to guide its use. With respect to activity contexts, there are a number of interesting scientific and technical avenues to explore.

Linguistic analyses

One step we would like to take is to assess an activity's *quality.* Rather than considering only a word's overall distribution across activities, does a word's use in certain activities carry more value than others? Intuitively, one might imagine that `storytime` is a higher value activity than others with respect to word learning. Along these lines, can we trace the origins of words in the child's vocabulary to different activities? Do new activities appear, perhaps after the introduction of a new game or toy, and do these give rise to a new set of word births?

Our analysis focused on caregiver word use *across* activities, but how a word is used *in* an activity is clearly also of interest. The work by Jerome Bruner that influenced our study of activity contexts focused on how language use is structured by a particular activity. Different approaches might be taken to studying a word's use during an activity. One potentially fruitful approach is through manual, qualitative analysis. Following Bruner, identifying the "deep structure" of an activity and the social roles, action sequences, and the link to language may then reveal something about the quality of different activities, mentioned above, and also provide the basis for looking at behavioral changes in an activity over time.

While manual, qualitative analyses may be powerful, they can also run up against the difficulty of working with a large dataset. Although not every analysis need leverage the full scale of the data, part of our interest is in discovering not only what can be learned from a full scale analysis, but also how it might be performed. Hence, operationalizing relevant measures of activity structure and "within activity" language use, and linking these measures to word births is a natural next step. One straightforward possibility is to look at a word's predictability in an activity. For example, given that an activity is taking place, how likely is a word to be used? Or if an activity's structure has been characterized, does the word hold a consistent relationship to some aspect of the activity, such as the word "peekaboo" in the game of peekaboo studied by Bruner?

Advanced techniques

We have claimed that activities can be modeled as latent variables that are reflected as distributions across location, time, participants and words. Instead of using LDA, a model that fully incorporates all of these modalities may yield better, more interpretable activity contexts. One model that we have considered is Dirichlet Multinomial Regression (DMR) (Mimno and McCallum, 2008). In this model, arbitrary document-level attributes can influence the prior topic distribution. In our case, this might be information about location, time and participants. However, at its core this is still a word-distribution model, but a better model of activity might build spatial, temporal, participant and word distributions into the topic itself. For example, the `making_coffee` activity has a very clear temporal and spatial distribution but little associated speech and may be poorly represented by topic models such as LDA and DMR.

While an activity may be clearly recognizable, as with any categorization scheme there can be different levels of granularity in the activity definitions. For example, an activity such as `reading` may be composed of more specific activities such as `reading_cat-in-hat`, `reading_bambi` and so on. This suggests a hierarchical model, but as model complexity increases the computational requirements and interpretability may become problematic.

Before exploring more sophisticated models for activity contexts, an immediate next step is to extract better "documents". For our work, documents were at 10 minute boundaries irrespective of changes in the participants' location, long gaps between utterances, or other possible discourse boundaries. Clearly activities can start and stop at arbitrary times, and allowing for flexible document endpoints may yield more coherent activity contexts.

7.3.3 Mutual adaptation

In investigating the role that environmental variables play in early word learning, we have attempted to distill some aspect of the child's experience with a word into a single number, such as the frequency or recurrence value of the word. The implicit assumption is that this number characterizes something about this word in the child's experience that may be relevant to its acquisition. But just as the child develops over the course of the study, so too does the environment. The key elements of the child's "environment" are his caregivers, who are also sensitive to the child's behavior and adapting to his needs and abilities.

Distilling aspects of a word in the environment into a single number, such as frequency, suggests a dynamic learner decoupled from a static environment, which is at odds with the underlying philosophy of this thesis. Ideas such as Vygotsky's (1986) Zone of Proximal Development and Bruner's (1983) observation that mothers guide children toward their potential suggest a tight coupling, an overall system of mutual adaptation between child and caregivers. This is a rich, if challenging, area for future research. In our own work, we have found evidence for caregivers tuning to the child's lexical learning. Caregiver utterance lengths containing a target word decrease until the point of the word birth, when they begin to increase again. As (Brent and Siskind, 2001) has shown, and our own work confirms (Vosoughi et al., 2010), words embedded in shorter utterances are learned earlier. In what other ways does the environment adjust to the child's competence and facilitate learning?

7.3.4 What's missing?

The four environmental variables we have considered are all predictive of word learning to varying degrees. But even if there were no overlap in the information they contributed, there would still be a great deal of unexplained variance in the child's word learning. How many more variables would be required to fully explain the child's language development? The extent to which external forces can shape human behavior and development is a fundamental question at the heart of philosophy, ethics, psychology, and other areas of thought. In this work, we have sometimes thought of environmental structures as "drivers" of learning, and other times as "footholds" for learning processes to cling to. But the child's learning also seems driven by other forces – his interests and preferences, to name a few. The child's vocabulary related to cars, trucks, and vehicles of all kinds was highly developed, a clear reflection of his interests. It seems likely that the idiosyncrasies and individual differences in children's vocabularies will derive not just from their differing environments but also from their individual personalities.

Nevertheless, we believe much more can be done in characterizing the environment and modeling word learning. Studying the adaptive feedback loops mentioned above, better characterizing the child's participation in activities, and modeling the learning process with more sophisticated models are all exciting directions for future work.

7.4 Final words

This thesis documented the recording, annotation, and the first analysis of the data collected for the Human Speechome Project. From the largest-ever record of a child's early development, we have followed along as he took his first steps, uttered his first words, socialized with his family, and marveled aloud at the wondrous world around him.

Our study is based on one child, and in contrast to analyses based on samples from many children care must be taken when making generalizations. But in a detailed study of one child, it is clear that the full richness of early experience contributes to word learning and is not easily captured with only a few samples. And just as every child's experience is different, there *are* common developmental patterns. Dan Slobin, in reflecting on his experience working with Roger Brown in the early 1960s, said "...the fact that the findings of those first years now seem familiar attests to the *cumulative* nature of child language study. We *have* learned something about the acquisition of English, and similar patterns have repeated themselves in enough children, in enough studies, to give us a feeling of secure knowledge." (Slobin, 1988, p. 11). We hope that our study has contributed to this knowledge, even as we hope it has led us deeper into the unknown territory of human language.

Appendix A

The bias of entropy and KL-divergence estimates for sampled multinomial distributions

Estimating the entropy of a multinomial distribution by using the unbiased, maximum likelihood estimate of the multinomial parameters leads to a biased result. Consequently, KL-divergence (relative entropy) estimates are also biased. As the number of samples n used to estimate the multinomial distribution increases, entropy and KL-divergence bias decreases. But for small n, or when comparing distributions estimated from different numbers of samples, this bias may introduce significant artifacts. The following derivations demonstrate how the expected entropy and KL-divergence vary with n, resulting in a straightforward relationship that depends only on n and B, the number of bins in the multinomial.

The derivation proceeds as follows. First, a bound on the bias of the entropy estimate is derived. Then, the cross-entropy is shown to be unbiased. By rewriting the KL-divergence in terms of the entropy and cross-entropy, a bound on the KL-divergence bias is obtained. Simulations confirm the results.

A.1 Expected entropy lower bound

Consider a multinomial distribution p with B bins, and estimates of p obtained by sampling. Let p_n be the distribution obtained by taking n samples from p. What is the *expected* entropy of p_n as a function of n?

We should expect that for $n = 1$ samples, all the probability mass will be in one particular bin of p_n and the entropy should be 0. As $n \to \infty$ we expect $p_n \to p$ and $H(p_n) \to H(p)$. This derivation explores this relationship.

We want to know the expected value of the entropy of p_n,

$$\begin{aligned} \mathrm{E}\left[H(p_n)\right] &= -\mathrm{E}\left[\sum_{i=1}^{B} p_{n,i} \log p_{n,i}\right] \\ &= -\sum_{i=1}^{B} \mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \end{aligned} \tag{A.1}$$

We now consider only the expected value term in (A.1). The maximum likelihood estimate of p from n samples is $p_{n,i} = \frac{x_i}{n}$ for all $i = 1 \ldots B$. Thus, for a particular bin i we have

$$\begin{aligned} \mathrm{E}\left[p_{n,i} \log p_{n,i}\right] &= \mathrm{E}\left[\frac{x_i}{n} \log \frac{x_i}{n}\right] \\ &= \sum_{k=0}^{n} \Pr(x_i = k) \frac{k}{n} \log \frac{k}{n} \end{aligned} \tag{A.2}$$

Assuming the samples are iid p, then the expected number of samples in bin i can be calculated as

$$
\begin{aligned}
&= \sum_{k=0}^{n} \binom{n}{k} p_i^k \left(1-p_i\right)^{n-k} \frac{k}{n} \log \frac{k}{n} \\
&= \frac{1}{n} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} \quad p_i^k \left(1-p_i\right)^{n-k} k \log \frac{k}{n} \\
&= \frac{1}{n} \sum_{k=1}^{n} \frac{n!}{(n-k)!k!} \quad p_i^k \left(1-p_i\right)^{n-k} k \log \frac{k}{n}
\end{aligned}
$$

Note that in this equation, p_i is the *true* probability of bin i in p rather than the estimated value. Also note that the last line above results from the fact that when $k = 0$, the summand is zero. Next we simplify $\frac{k}{k!}$, pull an n out of $n!$, and pull a p_i into the front of the sum obtaining

$$
= p_i \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} \quad p_i^{k-1} \left(1-p_i\right)^{n-k} \log \frac{k}{n}
$$

Now, let $j = k - 1$, $m = n - 1$ and apply Jensen's inequality for the following derivation:

$$
\begin{aligned}
&= p_i \sum_{j=0}^{m} \frac{m!}{(m-j)!j!} \quad p_i^j \left(1-p_i\right)^{m-j} \log \frac{j+1}{m+1} \\
&= p_i \sum_{j=0}^{m} \Pr(x_i = j) \log \frac{j+1}{m+1} && \text{(A.3)} \\
&\le p_i \log \sum_{j=0}^{m} \Pr(x_i = j) \frac{j+1}{m+1} && \text{(A.4)} \\
&= p_i \log \frac{mp_i + 1}{m+1} && \text{(A.5)} \\
&= p_i \log \frac{(n-1)p_i + 1}{n} && \text{(A.6)} \\
&= p_i \log \left(p_i + \frac{1-p_i}{n} \right) && \text{(A.7)}
\end{aligned}
$$

We obtain (A.4) from (A.3) using Jensen's inequality for the concave function $\log(\cdot)$. Equation (A.5) is just the expected value of the function $j+1$ for the binomial distribution. Substituting n back in for $m+1$ yields (A.6) which simplifies to (A.7).

Putting all this back together, we have

$$\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] \leq p_i \log\left(p_i + \frac{1-p_i}{n}\right) \tag{A.8}$$

$$-\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] \geq -p_i \log\left(p_i + \frac{1-p_i}{n}\right) \tag{A.9}$$

Recalling equation (A.1), and replacing the expectation with our lower bound we have

$$\mathrm{E}\left[H(p_n)\right] = \sum_{i=1}^{B} -\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] \tag{A.10}$$

$$\geq -\sum_{i=1}^{B} p_i \log\left(p_i + \frac{1-p_i}{n}\right) \tag{A.11}$$

Note that intuitively, as $n \to \infty$, $\log\left(p_i + \frac{1-p_i}{n}\right) \to \log p_i$ in equation (A.11) yielding the true entropy. When $n=1$, $\log\left(p_i + \frac{1-p_i}{n}\right) = \log(1) = 0$ implying $H(p_1) = 0$ as expected. Moreover, for each i, $-\log\left(p_i + \frac{1-p_i}{n}\right) < -\log p_i$ and therefore contributes a smaller factor to the total entropy, implying that for small n the expected entropy is smaller.

Equation (A.11) can also be written as

$$\begin{aligned} &\geq -\sum_{i=1}^{B} p_i \log\left(p_i\left(1 + \frac{1-p_i}{np_i}\right)\right) \\ &= -\sum_{i=1}^{B} p_i \log p_i - \sum_{i=1}^{B} p_i \log\left(1 + \frac{1-p_i}{np_i}\right) \\ &= H(p) - \sum_{i=1}^{B} p_i \log\left(1 + \frac{1-p_i}{np_i}\right) \end{aligned} \tag{A.12}$$

Equation (A.12) is useful because it isolates the component that varies with n. If $c_i = \frac{1-p_i}{p_i} \leq n$ then $x_i = c_i/n \leq 1$ and the Taylor expansion to $\log(1+x)$ can be applied. The Taylor series of $\log(1+x)$ for $-1 < x \leq 1$ is $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$ and so applying for $x = c_i/n$ we have

$$\log(1 + c_i/n) = \frac{c_i}{n} - \frac{c_i^2}{2n^2} + \frac{c_i^3}{3n^3} - \frac{c_i^4}{4n^4} + \dots$$

Plugging this into the summation in (A.12) yields the first line below, and in the second line we apply the fact that $p_i c_i^k = (1 - p_i) c_i^{k-1}$ to get

$$\begin{aligned}
&= -\sum_{i=1}^{B} p_i \left(\frac{c_i}{n} - \frac{c_i^2}{2n^2} + \frac{c_i^3}{3n^3} - \frac{c_i^4}{4n^4} + \dots \right) \\
&= -\sum_{i=1}^{B} \frac{1 - p_i}{n} + \sum_{i=1}^{B} \frac{(1 - p_i) c_i}{2n^2} - \sum_{i=1}^{B} \frac{(1 - p_i) c_i^2}{3n^3} \dots \\
&\approx -\frac{B - 1}{n} \qquad \text{(A.13)}
\end{aligned}$$

with the approximation improving as n increases. Also note that this preserves the bound, since including the first odd number of terms of the Taylor series (ie. 1,3,5,... terms) is always greater than the whole series. Therefore, we can write an alternative bound on the expected entropy as

$$\mathrm{E}\left[H(p_n)\right] \geq H(p) - \frac{B - 1}{n} \qquad \text{(A.14)}$$

This shows that the lower bound approaches the true entropy with $\frac{1}{n}$, a property that will come into play later for the KL-divergence.

A.1.1 Expected entropy upper bound

The lower bound of the expected entropy converges to the true entropy of the distribution as $n \to \infty$. This is not surprising, as $p_n \to p$. Nevertheless, we can also find an upper bound on the expected entropy to better understand how it varies with n.

Starting with equation (A.2), we have

$$\begin{aligned}
\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] &= \sum_{k=0}^{n} \Pr(x_i = k)\frac{k}{n}\log\frac{k}{n} \\
&= \sum_{k=0}^{n} \Pr(x_i = k)\frac{k}{n}\log\frac{k\cdot\Pr(x_i = k)}{n\cdot\Pr(x_i = k)}
\end{aligned}$$

Let $a_k = \Pr(x = k)\frac{k}{n}$ and $b_k = \Pr(x = k)$. Then rewriting, and applying the log-sum inequality yields

$$\begin{aligned}
&= \sum_{k=0}^{n} a_k \log\frac{a_k}{b_k} \\
&\geq \left(\sum_{k=0}^{n} a_k\right)\frac{\sum_{k=0}^{n} a_k}{\sum_{k=0}^{n} b_k} \\
&= p_i \log p_i
\end{aligned}$$

since $\sum_{k=0}^{n} a_k = p$ and $\sum_{k=0}^{n} b_k = 1$. Therefore, $\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] \geq p_i \log p_i$ or equivalently, $-\mathrm{E}\left[p_{n,i}\log p_{n,i}\right] \leq -p_i \log p_i$. Putting this bound on equation (A.2) back in for equation (A.1) gives

$$\begin{aligned}
\mathrm{E}\left[H(p_n)\right] &\leq -\sum_{i=1}^{B} p_i \log p_i \\
&= H(p)
\end{aligned}$$

In other words, the expected entropy of the sampled distribution obtained after n samples is upper bounded by the entropy of the true distribution.

It would be nice to find a tighter upper bound on the expected entropy, namely, one that varies with n. One crude way to show that the upper bound increases with n is to find the maximum entropy p_n for each n. The maximum entropy p_n would be one where each sample lands in a new bin, and for $n \leq B$ we have $p_i = \log 1/n$. However, this is not a very satisfying upper bound. It may be more fruitful to focus on probabilistic bounds, that may instead take the variance into account.

A.2 Expected cross entropy

The cross entropy between q and p, here denoted as $H(q,p) = -\sum_i q_i \log p_i$, can be thought of as the cost in bits of encoding q using a code for p. Suppose we have q_n – the distribution obtained by taking n samples from q. Then what is the expected cross entropy $H(q_n, p)$?

Similar to the expected entropy calculation, we seek

$$\begin{aligned} \mathrm{E}\left[H(q_n,p)\right] &= -\mathrm{E}\left[\sum_{i=1}^{B} q_{n,i} \log p_i\right] \\ &= -\sum_{i=1}^{B} \mathrm{E}\left[q_{n,i} \log p_i\right] \\ &= -\sum_{i=1}^{B} \mathrm{E}\left[\frac{x_i}{n} \log p_i\right] \end{aligned}$$

The expected value $E\left[x_i\right]$ above is just the expected count for bin i under the true probability distribution q. Thus, $E\left[x_i\right] = nq_i$ and

$$\begin{aligned} -\sum_{i=1}^{B} \frac{1}{n} \log p_i \; \mathrm{E}\left[x_i\right] &= -\sum_{i=1}^{B} q_i \log p_i \\ &= H(q,p) \end{aligned} \tag{A.15}$$

So the expected cross entropy $\mathrm{E}\left[H(q_n,p)\right]$ is just the true cross entropy $H(q,p)$. Note that $H(p,p) = H(p)$, and in the special case where $q_n = p_n$ we have that $\mathrm{E}\left[H(p_n,p)\right] = H(p)$.

A.3 Expected KL-divergence upper bound

The KL-divergence between distributions q and p is written as $D(q||p) = \sum_i q_i \log \frac{q_i}{p_i}$. This can be rewritten as

$$\begin{aligned} D(q||p) &= \sum_i q_i \log q_i - \sum_i q_i \log p_i && \text{(A.16)} \\ &= H(q,p) - H(q) && \text{(A.17)} \end{aligned}$$

For all q and p of the same dimension, $D(q||p) \geq 0$ with equality iff $q = p$. The KL-divergence can be thought of as the additional bits required to encode q using a code for p rather than the code for q. What is the expected KL-divergence of a sampled distribution p_n to the true distribution p? Using the results from the previous sections, we have

$$\begin{aligned} \mathrm{E}\left[D(p_n||p)\right] &= \mathrm{E}\left[H(p_n,p) - H(p_n)\right] \\ &= \mathrm{E}\left[H(p_n,p)\right] - \mathrm{E}\left[H(p_n)\right] \\ &= H(p) - \mathrm{E}\left[H(p_n)\right] \\ &\leq H(p) + \sum_{i=1}^{B} p_i \log\left(p_i + \frac{1-p_i}{n}\right) && \text{(A.18)} \end{aligned}$$

If we instead write the KL-divergence bound using equation (A.14) we have

$$\begin{aligned} \mathrm{E}\left[D(p_n||p)\right] &\leq H(p) - H(p) + \frac{B-1}{n} \\ &= \frac{B-1}{n} && \text{(A.19)} \end{aligned}$$

For comparing a sampled distribution q_n against p, we have

$$\begin{aligned} \mathrm{E}\left[D(q_n||p)\right] &\leq H(q,p) - H(q) + \frac{B-1}{n} \\ &= D(q||p) + \frac{B-1}{n} \end{aligned}$$

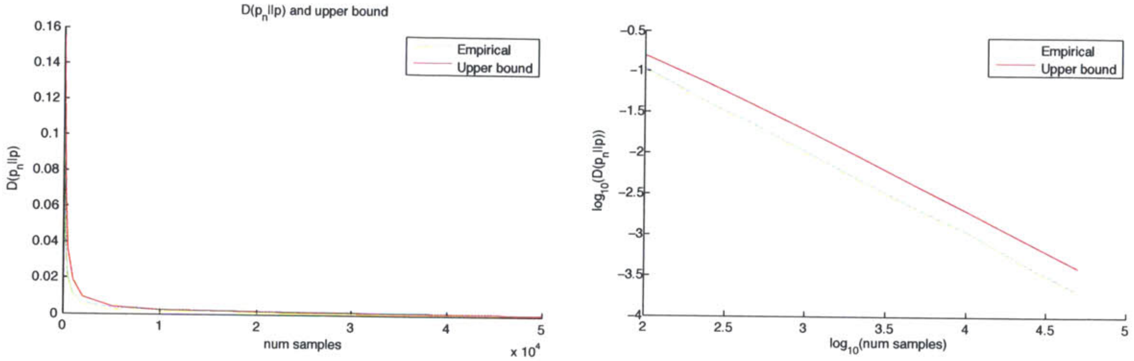


Figure A-1: KL-divergence for p_n sampled from p for various n, and the upper bound on the expected value.

In other words, for a given number of samples n, we expect the sampled KL-divergence to be within a certain range of the true KL-divergence, depending on n.

We tested this upper bound by taking n samples of p to obtain p_n, computing $D(p_n||p)$, and repeating this many times for each n. The average of $D(p_n||p)$ is an estimate of the expected KL-divergence for n. We then computed the upper bound using equation (A.18) and plotted this as well. The simulation results are shown in figure A-1.

Bibliography

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

Barabási, A. (2010). *Bursts: The hidden pattern behind everything we do.* EP Dutton.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Blei, D. (2011). Introduction to probabilistic topic models. *Communications of the ACM*, pages 1–16.

Blei, D. (2012). http://www.cs.princeton.edu/˜blei/lda-c/index.html.

Blei, D. and Lafferty, J. (2006a). Correlated topic models. In *Advances in Neural Information Processing Systems.*

Blei, D. and Lafferty, J. (2009). Topic models. In Srivastava, A. and Sahami, M., editors, *Text mining: Classification, clustering, and applications*, CRC Data Mining and Knowledge Discovery Series, page 71. Chapman & Hall.

Blei, D. M. and Lafferty, J. D. (2006b). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 113–120.

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

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax*, volume 154. Mouton, The Hague.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* The MIT Press.

Bowerman, M. (1978). The acquisition of word meaning: An investigation into some current conflicts. *The development of communication*, pages 263–287.

Braunwald, S. (1978). Context, word and meaning: Toward a communicational analysis of lexical acquisition. *Action, gesture and symbol: The emergence of language*, pages 485–527.

Brent, M. and Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:33–44.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design.* Harvard University Press.

Brown, R. (1973). *A first language: The early stages.* Harvard University Press.

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

Bruner, J. (1985). The role of interaction formats in language acquisition. In Forgas, J. P., editor, *Language and Social Situations*, chapter 2, pages 31–46. Springer-Verlag.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* The MIT Press.

Chomsky, N. (2005). *Rules and representations.* Columbia University Press.

Clark, E. (1995). *The lexicon in acquisition*, volume 65. Cambridge University Press.

Clauset, A., Shalizi, C., and Newman, M. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Comrie, B. (2000). From potential to realization: An episode in the origin of language. *Linguistics*, 38(5):989–1004.

Cover, T. and Thomas, J. (2006). *Elements of information theory.* John Wiley & Sons.

DeCamp, P. (2007). HeadLock: Wide-range head pose estimation for low resolution video. Master's thesis, Massachusetts Institute of Technology.

Dromi, E. (1987). *Early Lexical Development.* Cambridge University Press.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D. J., Pethick, S., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59:1–185.

Fiscus, J. (2007). Speech recognition scoring toolkit ver. 2.3 (sctk).

Ganger, J. and Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, 40(4):621.

Gentner, D. (2006). Why verbs are hard to learn. *Action meets word: How children learn verbs*, pages 544–564.

Gentner, D. and Boroditsky, L. (2001). Individuation, relativity, and early word learning. In Bowerman, M. and Levinson, S. C., editors, *Language acquisition and conceptual development*, page 215. Cambridge University Press.

Goldfield, B. A. and Reznick, S. J. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, 17(1):171–183.

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

Gopnik, A. and Meltzoff, A. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development*, 58:1523–1531.

Hart, B. and Risley, T. (1995). *Meaningful Differences in the Everyday Experience of Young American Children.* Brookes Publishing Company, Baltimore, MD.

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

Hurtado, N., Marchman, V., and Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, 11(6):F31–F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., and Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27(1236-248).

Jusczyk, P. (1997). *The discovery of spoken language.* The MIT press.

Kubat, R., DeCamp, P., Roy, B., and Roy, D. (2007). TotalRecall: Visualization and semi-automatic annotation of very large audio-visual corpora. In *ICMI.*

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3):299 – 321.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Locke, J. (1689/2008). *An Essay Concerning Human Understanding.* Oxford University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Lawrence Erlbaum Associates, third edition.

Mayor, J. and Plunkett, K. (2010). Vocabulary spurt: Are infants full of Zipf? In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 836–841.

Mayor, J. and Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, 14(4):769–785.

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

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, 317:631.

Medina, T., Snedeker, J., Trueswell, J., and Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014.

Miller, M. (2011). Semantic spaces: Behavior, language and word learning in the human speechome corpus. Master's thesis, Massachusetts Institute of Technology.

Mimno, D. and McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 411–418, Corvallis, Oregon. AUAI Press.

Mitchell, C. C. and McMurray, B. (2008). A stochastic model for the vocabulary explosion. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, pages 1–135.

Peters, A. M. (1983). *The Units of Language Acquisition*. Cambridge University Press.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

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

Reidsma, D., Hofs, D., and Jovanović, N. (2005). Designing focused and efficient annotation tools. In *Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands.

Rice, J. (2007). *Mathematical statistics and data analysis*. Duxbury press, 3rd edition.

Roy, B. C. (2007). Human-machine collaboration for rapid speech transcription. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

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

Roy, B. C., Frank, M. C., and Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Roy, B. C. and 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., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., and Gorniak, P. (2006). The human speechome project. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pages 2059–2064, Mahwah, NJ. Lawrence Erlbaum Associates.

Roy, D. K. (1999). *Learning Words from Sights and Sounds: A Computational Model.* PhD thesis, Massachusetts Institute of Technology.

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

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928.

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

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39 – 91. ¡ce:title¿Compositional Language Acquisition¡/ce:title¿.

Slobin, D. (1988). From the Garden of Eden to the Tower of Babel. In Kessel, F. S., editor, *The Development of Language and Language Researchers: Essays in Honor of Roger Brown*, pages 9–22. Lawrence Erlbaum Associates.

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a word learner: A debate on lexical acquisition*, chapter 3, pages 51–80. Oxford University Press.

Snow, C. E. (1988). The last word: Questions about the emerging lexicon. In Smith, M. D. and Locke, J. L., editors, *The emergent lexicon: The child's development of a linguistic vocabulary*, chapter 11, pages 341–353. Academic Press.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition*, 50(1-3):431–45.

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1):89–96.

St. Augustine (398/1961). *The Confessions of St. Augustine.* Random House.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tomasello, M. and Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31(1):101–121.

Trueswell, J., Medina, T., Hafri, A., and Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1):126–156.

Tsourides, K. (2010). Visually Grounded Virtual Accelerometers: A Longitudinal Video Investigation of Dyadic Bodily Dynamics around the time of Word Acquisition. Master's thesis, Massachusetts Institute of Technology.

Vosoughi, S. (2010). Interactions of caregiver speech and early word learning in the Speechome Corpus: Computational Explorations. Master's thesis, Massachusetts Institute of Technology.

Vosoughi, S. (2012). Personal communication.

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

Vosoughi, S. and Roy, D. (2012). An automatic child-directed speech detector for the study of child language development. In *Proceedings of Interspeech*, Portland, Oregon.

Vygotsky, L. (1986). *Thought and Language.* The MIT Press.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Technical Report 139, Sun Microsystems.

Weisleder, A. (2011). Personal communication.

Weisleder, A. and Fernald, A. (under review). Talking to children matters: Early language experience strengthens processing and builds vocabulary.

Windsor, J., Glaze, L., Koga, S., et al. (2007). Language acquisition with limited input: Romanian institution and foster care. *Journal of Speech, Language and Hearing Research*, 50(5):1365.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Series in Data Management Systems. Morgan Kaufmann, second edition.

Wittgenstein, L. (1953/2009). *Philosophical Investigations.* Wiley-Blackwell, 4th edition.

Wittgenstein, L. (1965). *The Blue and Brown Books.* Harper Perennial.

Yu, C. and Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414.

Zeanah, C., Nelson, C., Fox, N., Smyke, A., Marshall, P., Parker, S., Koga, S., et al. (2003). Designing research to study the effects of institutionalization on brain and behavioral development: The Bucharest Early Intervention Project. *Development and Psychopathology*, 15(4):885–907.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort.* Addison-Wesley.