Interactions of caregiver speech and early word learning in the Speechome Corpus: Computational Explorations

by

Soroush Vosoughi

B.S., 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

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author_____

Program in Media Arts and Sciences
Aug 10, 2010

Certified by_____

Deb K. Roy
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by_____

Pattie Maes
Associate Academic Head
Program in Media Arts and Sciences

Interactions of caregiver speech and early word learning in the Speechome Corpus: Computational Explorations

by

Soroush Vosoughi

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on Aug 10, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

How do characteristics of caregiver speech contribute to a child's early word learning? What is the relationship between a child's language development and caregivers' speech? Motivated by these general questions, this thesis comprises a series of computational studies on the fined-grained interactions of caregiver speech and one child's early linguistic development, using the naturalistic, high-density longitudinal corpus collected for the Human Speechome Project. The child's first productive use of a word was observed at about 11 months, totaling 517 words by his second birthday. Why did he learn those 517 words at the precise ages that he did? To address this specific question, we examined the relationship of the child's vocabulary growth to prosodic and distributional features of the naturally occurring caregiver speech to which the child was exposed. We measured fundamental frequency, intensity, phoneme duration, word usage frequency, word recurrence and mean length of utterances (MLU) for over one million words of caregivers' speech.

We found significant correlations between all 6 variables and the child's age of acquisition (AoA) for individual words, with the best linear combination of these variables producing a correlation of $r = -.55 (p < .001)$. We then used these variables to obtain a model of word acquisition as a function of caregiver input speech. This model was able to accurately predict the AoA of individual words within 55 days of their true AoA. We next looked at the temporal relationships between caregivers' speech and the child's lexical development. This was done by generating time-series for each variables for each caregiver, for each word. These time-series were then time-aligned by AoA. This analysis allowed us to see whether there is a consistent change in caregiver behavior for each of the six variables before and after the AoA of individual words.

The six variables in caregiver speech all showed significant temporal relationships with the child's lexical development, suggesting that caregivers tune the prosodic and distributional characteristics of their speech to the linguistic ability of the child. This tuning behavior involves the caregivers progressively shortening their utterance lengths, becoming more redundant and exaggerating prosody more when uttering particular words as the child gets

closer to the AoA of those words and reversing this trend as the child moves beyond the AoA. This "tuning" behavior was remarkably consistent across caregivers and variables, all following a very similar pattern. We found significant correlations between the patterns of change in caregiver behavior for each of the 6 variables and the AoA for individual words, with their best linear combination producing a correlation of $r = -.91 (p < .001)$. Though the underlying cause of this strong correlation will require further study, it provides evidence of a new kind for fine-grained adaptive behavior by the caregivers in the context of child language development.

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

Interactions of caregiver speech and early word learning in the Speechome Corpus: Computational Explorations

by

Soroush Vosoughi

The following people served as readers for this thesis:

Thesis Reader_____

Dr. John Makhoul
Chief Scientist
BBN Technologies

Thesis Reader_____

Dr. Rochelle Newman
Associate Professor of Hearing and Speech Sciences
University of Maryland

Acknowledgements

I have never been good with words, which is why I find myself in such a delicate conundrum to give everyone the thanks they deserve.

First and foremost I would like thank my adviser, Prof. Deb Roy for his advise, continued support and encouragement throughout my years at MIT. I also would like to thank my readers, Dr. Makhoul and Prof. Newman for providing feedback on this thesis on such a short notice. Many thanks to Michael Frank for his valuable advice and insight.

Thanks to all the members of the Cogmac group, particularly Brandon Roy, whose friendship and advice contributed to a very positive work environment.

A big thank you goes to all of my friends at MIT, graduate and undergraduate, who have made my years at MIT the best years of my life thus far. Particularly, I would like to thank Alice for her everlasting and contagious optimism.

A special thank you goes to Kai-yuh Hsiao, who was without a doubt the best UROP supervisor anyone could ever ask for. There is absolutely no doubt in my mind that without him I would not be where I am today. I probably learned more from Kai-yuh than I did from all my classes at MIT.

Last, but certainly not least, I would like to thank my dad, mom, brother and sister, whose continued support and encouragement so far has seen me through six years of MIT.

Soroush Vosoughi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Human Speechome Project

The Human Speechome Project (HSP) [30] was launched in 2005 to study early language development through analysis of audio and video recordings of the first two to three years of one child's life. The home of a family with a young child was outfitted with fourteen microphones and eleven omni-directional cameras at the time of birth of their first child. Audio was recorded from ceiling mounted boundary layer microphones at 16 bit resolution with a sampling rate of 48 KHz. Due to the unique acoustic properties of boundary layer microphones, most speech throughout the house including very quiet speech was captured with sufficient clarity to enable reliable transcription. Video was also recorded to capture non-linguistic context using high resolution fish-eye lens video cameras that provide a bird's-eye view of people, objects, and activity throughout the home. For more information about the recording infrastructure of the HSP please read The Human Speechome Project [30].

The Human Speechome project captures one child's development in tremendous depth. While this aspect of the project limits conclusions about general aspects of language development (as with previous longitudinal case studies [26, 33]), the dense sampling strategy affords many advantages over other corpora (eg. [15]). First, the HSP corpus is higher in density than other reported corpus, capturing an estimated 70% of the child's wakeful

experiences during the recording period. Second, since data were collected without specific theoretical assumptions or hypotheses, they can be reanalyzed in multiple ways from different theoretical perspectives. Finally, since high resolution video was also collected the role of non-linguistic context can also be studied (though in the current study we restrict our analysis to aspects of speech input).

1.2 Motivations

How do characteristics of caregiver speech contribute to a childs early word learning? What are the mechanisms underlying child language acquisition? What is the relationship between a childs language development and caregivers' speech? Are there correlations between the input the child receives and his lexical development? Are aspects of children's input predictive of the child's later lexical development? These are the main questions that motivate this thesis. Answering these questions could help us understand the nature of language learning strategies and mechanisms used by children.

Even though there is a large literature of research in this area, our overall understanding of the interactions of caregiver speech and early word learning remains limited by the lack of appropriate data. The dense, longitudinal and naturalistic nature of the Speechome corpus, allows us for the first time to study the fined-grained(at the level of single words) relationships between caregiver speech and early word learning.

As a general goal, we want to predict the child's language outcome based on his linguistic interactions with the caregivers. We are also interested in understanding whether the relationship between the child's lexical acquisition and caregiver speech is bidirectional. That is, whether the caregivers' linguistic behavior change based on the lexical development of the child.

This has obvious clinical and theoretical utility. From the clinical perspective, such findings can in the future(when done on a large enough sample) help us identify kids that are at risk for language problems. From a theoretical perspective, if certain signals in caregiver

speech are found to be predictive of age of acquisition of words, the assumption is that those signals are somehow important for acquisition of words. Thus, such findings can help us sort out which factors are really important for the child's process of language learning. Moreover, such findings might tell us something about the factors and signals in caregiver speech that capture the child's attention. Basically, the point is that the actually findings from these studies are in some way less important than what they could imply about the language-learning process more generally [22].

The final motivation for this thesis is we hope by that understanding the nature of language learning strategies and mechanisms used by children, we can design and implement artificial language learning systems that use similar learning mechanisms to that of a child.

However, as mentioned, even though the unique nature of the Speechome corpus allows for fined-grained analysis of the child's early word learning, the fact that the Speechome corpus captures only one child's language development in some ways limits the conclusions about child language development that may be drawn from our analysis of this corpus.

1.3 Key Contributions

Here we provide a brief summary of the key contributions of this thesis. For detailed analysis and explanation of these contributions please read the entire thesis document. The major contributions of this thesis are:

- Development of independent and creative methods of data analysis. Including a collection of tools integrated into a software analysis environment for fast processing and analyzes of high density, longitudinal corpora such as the Speechome corpus.
- Exploration of the relationship between child word learning and six prosodic and distributional features(fundamental frequency, intensity, phoneme duration, usage frequency, recurrence and mean length of utterances) of the naturally occurring caregiver speech that the child was exposed to. Showing significant correlation between these

variables coded in caregiver speech and age of acquisition of words by the child. Using these variables to obtain and evaluate a highly predictive model of word acquisition as a function of caregiver input speech.

- Study of the fine-grained temporal relationships between caregivers' speech and the child's lexical development. Showing significant evidence of caregiver tuning for all of the 6 variables that we coded in caregiver speech. Using patterns of change in caregiver behavior to obtain and evaluate a model of word acquisition as a function of change in caregiver behavior for each of the 6 variables.

- Preliminary analysis of the difference between child directed speech(speech directed at the child) and child available speech(speech available to the child) in relation to our predictive models of child word acquisition.

- Development and evaluation of a fully automatic system capable of replicating (though rather poorly) all studies done in the thesis from raw audio in the corpus without any human processing or transcription.

- Development of an on-line prediction system capable of calculating the probability of words beings learned by the child in real time, while listening to the audio corpus.

1.4 Outline of the Thesis

The current chapter is an introduction to the Human Speechome Project and a brief overview of the motivations and contributions of this thesis. The remaining chapters are arranged as follows:

- Chapter 2 describes in great detail the HSP corpus and the tools that were developed to analyze this corpus. This chapter also provides detailed definitions for seven predictor variables- frequency, recurrence, mean length of utterance, duration, F0 and intensity- coded in caregiver speech and used for our analysis throughout this thesis.

- Chapter 3 explores in detail the correlations of the seven predictor variables (and their linear combinations) coded in caregiver speech with the age of acquisition of words by the child.
- Chapter 4 describes the thesis's first predictive model capable of predicting the age of acquisition of words by the child using the six predictive variables coded in caregiver speech.
- Chapter 5 highlights several limitations to the first predictive model, developed in Chapter 4.
- Chapter 6 addresses one of the main limitations mentioned in Chapter 5, the linear input-output aspect of the first predictive model. This problem is addressed by exploring the mutual influences between caregivers and the child, which show strong evidence of caregiver "tuning" with respect to the lexical development of the child.
- Chapter 7 describes the thesis's second predictive model capable of predicting the age of acquisition of words by the child. This model utilizes valleys that appear in the second derivatives of the mutual influence curves generated in Chapter 6 to predict the AoA of words. In order to automatically detect these valleys, we also developed and evaluated a valley detector in this chapter.
- Chapter 8, describes the development and evaluation of a child directed speech detector. We then use this detector to generate new predictive models based on child directed speech. We conclude by comparing the performance of the child directed speech trained models with our previous child available speech trained models.
- Chapter 9, oversees the development and evaluation of an automatic speech recognizer for the HSP corpus and use that to build a fully automatic predictive model. This model is capable of predicting the age of acquisition of words by the child by processing and analyzing raw audio from the HSP audio corpus. We then compare the performance of this model with our previous models.
- In Chapter 10, we develop an on-line predictive model. This model "listens" to caregiver speech from the HSP corpus in a chronological fashion(day by day). At the end

of each day it predicts the probability of the child having already acquired a word, for every word in the corpus.

- Chapter 11 reiterates the contributions of this thesis.
- Chapter 12 outlines plans for future work.

Chapter 2

Methods

This chapter describes in detail the datasets, tools and systems that were developed and used during the course of this thesis.

2.1 The Speechome Corpus

The dataset collected for the Human Speechome Project comprises more than 120,000 hours of audio and 90,000 hours of video. Most analysis depends on annotated data, however, so an effective annotation methodology is critical to the project's success. Brandon Roy has developed a semi-automated speech transcription system called BlitzScribe that facilitates fast and accurate speech transcription [28]. Automatic speech detection and segmentation algorithms identify speech segments, presenting them to a human transcriber in a simple user interface. This focuses human effort on the speech and leads to a highly efficient transcription process. Using BlitzScribe, transcribers were able to obtain an approximately five-fold performance gain at comparable accuracy to other tools.

Speaker identification algorithms are then applied to the transcribed audio segments, selecting from one of the four primary speakers (mother, father, nanny, and child) and producing a classification confidence score. Speaker annotation tools allow a human to review low

confidence segments and make corrections as necessary. Since identifying child directed speech(CDS) currently requires significant human effort, we operationalized the definition to refer to caregiver speech when the child is awake and close enough to hear. We refer to this as "child available speech" (CAS). Moreover, it is unclear whether CDS alone is used by children to learn language or if they also learn from "overheard" speech in CAS. We did later-on develop an automatic CDS detector for our corpus but did not have sufficient time to redo our analysis. We will go over this analysis in section 9 of the thesis where we also compare the effects of CAS and CDS on our analysis.

Given our interest in early word learning, the analysis done in this thesis focuses on the child's 9–24 month age range, and the corresponding subset of the corpus contains 4260 hours of 14-track audio, of which and estimated 1150 hours contain speech. Of the 488 days in this time range, recordings were made 444 of the days with a mean of 9.6 hours recorded per day. The current results are based on 218 fully transcribed days containing an average of 28,712 words per day of combined CAS and child speech, totaling 6.26 million words. It is estimated that the fully transcribed 9–24 month corpus will contain 12 million words. The long term goal is to fully annotate all speech in the corpus with transcriptions, speaker identity, and prosodic features.

Three limitations of the speech annotation process required us to filter the 3.87 million words of transcripts and only use a subset of the transcripts for the current analyses. First, roughly 2400,000 words belong to utterances marked by human transcribers as containing more than one speaker. In other words, about 39% of pause separated spoken utterances contain abutting or overlapping speech of two or more people, reflecting the realities of "speech in the wild". Since we cannot currently automatically distinguish the sources of this type of speech , we removed these utterances. Second, to reduce errors due to automatic speaker identification, we sorted utterances based on a confidence metric produced by the speaker identification algorithm and removed approximately the bottom 50% of utterances. Third, about 15% of the remaining utterances were deemed by human transcribers to be of insufficient clarity to transcribe reliably. After removing those utterances, we obtained the 1,031,201 word corpus used for all analyses in this paper.

2.2 Parallelized Infrastructure

The analysis methods developed in this thesis, when applied to the Speechome audio corpus, demand significant computing power. In order to be able to process and analyze our corpus in any reasonable amount of time, we developed a parallelized algorithm similar to map-reduce [36] that runs across four Intel quad-core machines. Each of these four machines can run 4 simultaneous client software (one for each core). All of these clients are blind to each other and are connected to a single host machine. Figure 2-1 shows the schematics of our parallelized infrastructure. Though we only used 16 cores to do the analysis in this thesis, the system is designed such that it can run across as many computers and cores that are available to the user.

The job of the host is to divide up every task into many independent parts and sends them to the clients to be processed. After a client is done processing the job the results are sent back to the host which then merges the results from the clients. The system is very robust in that the failure of any number of clients would not have any adverse effect on the system other than slowing it down. Any job that was passed to a client that has crashed would just be reassigned to another client at a later time. As long as there is at least one client still running the system is guaranteed to complete the task.

2.3 Predictor Variables' Definitions and Extraction Methods

In this section we will define in detail seven predictor variables coded in caregiver speech (plus the AoA) and describe how they were extracted from the HSP corpus. Figure 2-2 shows the general pipeline used to extract these variables from our speech and transcription files.

Three of the seven variables- duration, fundamental frequency and intensity- are proxies for prosodic emphasis. Previous studies have shown that infants are sensitive to the prosodic aspects of speech [16, 3, 2]. It has also been suggested that prosody is used by children for word segmentation [8, 7, 10, 12, 13, 18, 19, 20, 24, 23, 25] .

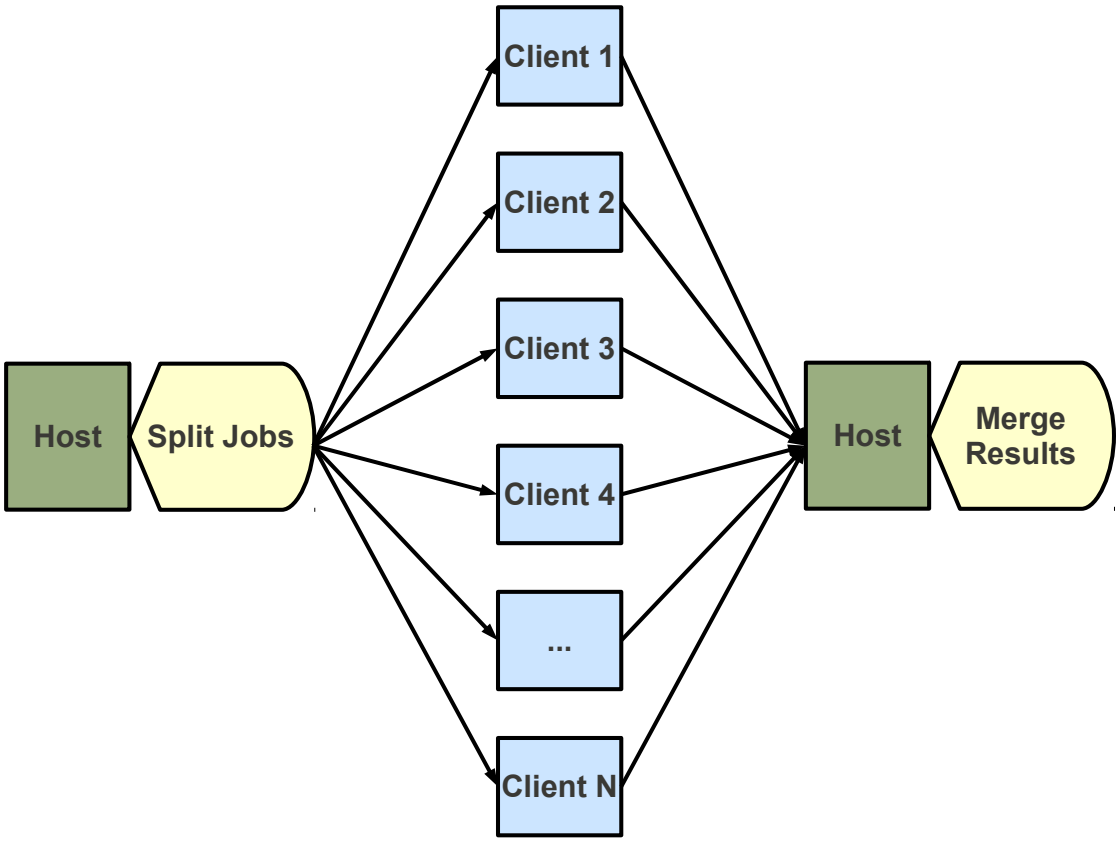


Figure 2-1: General design principle behind the parallelized infrastructure. Note that the only communication pipeline is between the host and the clients and so none of the clients are aware of each other.

The other four variables are: frequency, recurrence, mean length of utterances and time-of-day. Previous studies [11, 9] have shown at least one of these variables(frequency) to be correlated with age of acquisition of words by children.

Below we give the operational definition that we ended up using for age of acquisition and for each of the seven predictor variables that we use in our analysis. All variables for a particular word are computed using CAS up to the AoA for that word.

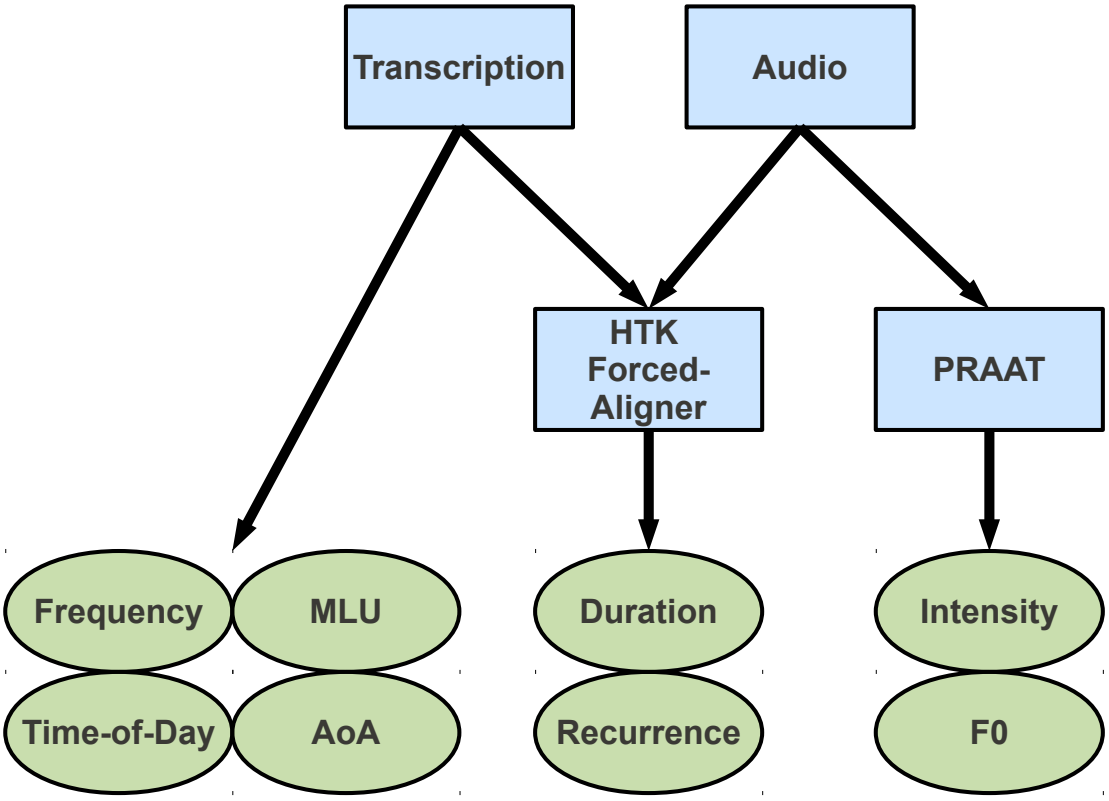


Figure 2-2: Schematic of the processing pipeline for outcome and predictor variables.

2.3.1 Age of Acquisition

We defined the AoA for a particular word as the first time in our transcripts that the child produced a word. Using this definition, the first word was acquired at nine months of age

with an observed productive vocabulary of 517 words by 24 months (though the actual productive vocabulary might be considerably larger when transcription is completed). In order to ensure reliable estimates for all predictors, we excluded those words from the child's vocabulary for which there were fewer than six caregiver utterances. This resulted in the exclusion of 56 of the child's 517 words, leaving 461 total words included in the current analysis.

2.3.2 Frequency

The frequency predictor measures the log of the count of word tokens in CAS up to the time of acquisition of the word divided by the period of time over which the count is made. Thus, this measure captures the average frequency over time of a word being used in CAS.

2.3.3 Recurrence

Distinct from frequency, recurrence measures the average repetition of a particular word in caregiver speech within a short window of time. Figure 2-3 highlights the difference between recurrence and frequency. As shown in Figure 2-3, W2 has the same frequency in both utterances, however its average recurrence (measured in an arbitrary window of time for this example) differs.

The window size parameter needed to be set to some constant. The window size could have been set to be anywhere from a few seconds to a few minutes. We wanted the window size that generates the greatest correlation between recurrence and AoA. Therefore, the window size was set by searching all possible window sizes from 1 to 600 seconds using our parallelized infrastructure. For each window size, we performed a univariate correlation analysis to calculate the correlation between recurrence at that window size and AoA. We then selected the window size which produced the largest correlation at (51 seconds). Figure 2-10(a) shows the correlation between AoA and recurrence at each window size.

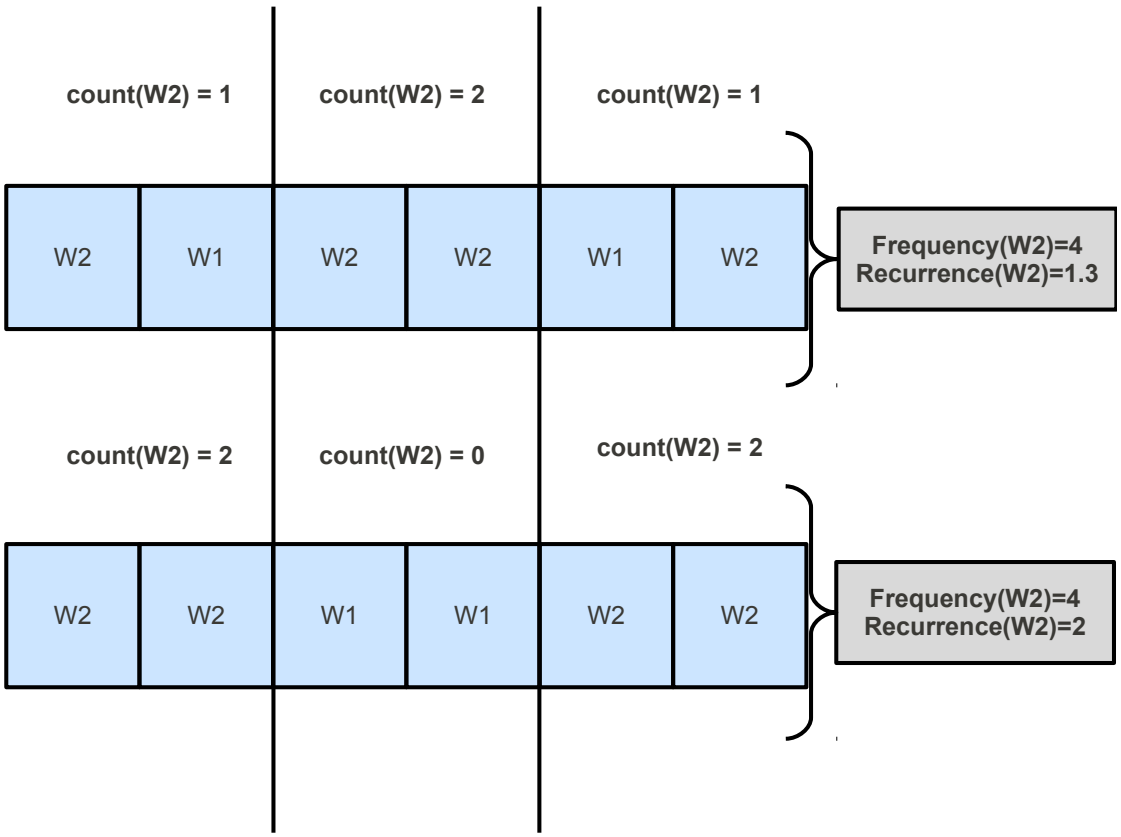


Figure 2-3: An example highlighting the difference between frequency and recurrence.

2.3.4 Mean Length of Utterances (MLU)

The MLU [17] predictor measures the mean utterance length of caregiver speech containing a particular word. Note that we report MLU based on the number of orthographic words(as opposed to morphemes) in each utterance. Figure 2-4 shows an example of how MLU is calculated.

In order to be consistent with the direction of correlation for other variables (a negative correlation with the AoA) we use 1/MLU as the predictor. From now on, whenever all mentions of MLU are to be treated as 1/MLU.

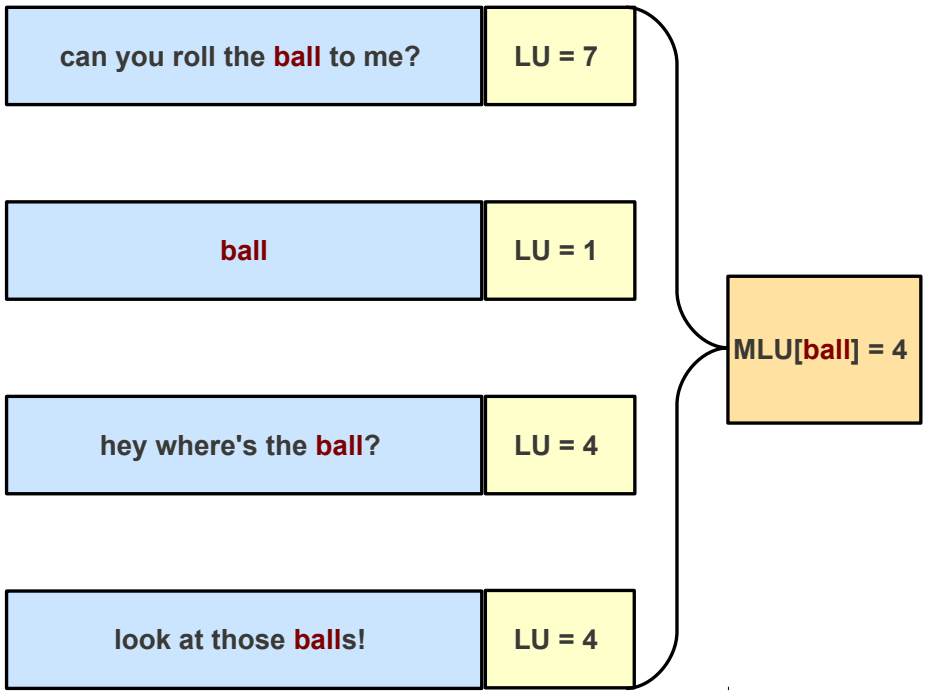


Figure 2-4: An example of how MLU is calculated.

2.3.5 Duration

The duration predictor is a standardized measure of word duration for each word [14]. We first extracted duration for all phoneme tokens in the corpus. We next converted these to

normalized units for each phoneme separately (via z-score), and then measured the mean standardized phoneme duration for the tokens of a particular word type. For example, a high score on this measure for the word "dog" would reflect that the phonemes that occurred in tokens of "dog" were often long relative to comparable phoneme sounds that appeared in other words. We grouped similar phonemes by converting transcripts to phonemes via the CMU pronunciation dictionary [35]. As with the recurrence variable, duration also has a parameter that needs to be set. We needed to know which class of phonemes to use when calculating duration. There were three possible choices: all phonemes, vowels only and sonorants only. We wanted the class of phonemes that generates the greatest correlation between duration and AoA. Therefore, we selected the class by trying all three classes. For each class, we performed a univariate correlation analysis to calculate the correlation between duration using that phoneme class and AoA. The class that produced the largest correlation was the phoneme class which was then selected. Figure 2-10(b) shows the correlation between AoA and duration for each phoneme class.

The extraction of phoneme duration for all phoneme tokens in the corpus was done automatically using a forced-aligner. Below is a description of how the forced-aligner works.

2.3.5.1 Forced-Aligner

A forced-aligner is almost identical to an automatic speech recognizer with one main difference, the force-aligner is given a transcription of what is being spoken in the audio data. Figure 2-5 shows the pipeline used by the forced-aligner. The aligner works by first converting the transcripts to phonemes vis the CMU pronunciation dictionary. It then aligns the transcribed data with the speech data, identifying which time segments in the speech data correspond to particular phonemes in the transcription data. Figure 2-6 shows a sample phoneme alignment for an utterance. The forced-aligner that was used in this thesis is from the Hidden Markov Model Toolkit(HTK) [38]. In order for the forced-aligner to be able to match phoneme in the audio to the phonemes in the transcript, it needs to have an acoustic model for each of the phonemes in the transcript. In the next section we will describe how we obtained our acoustic models.

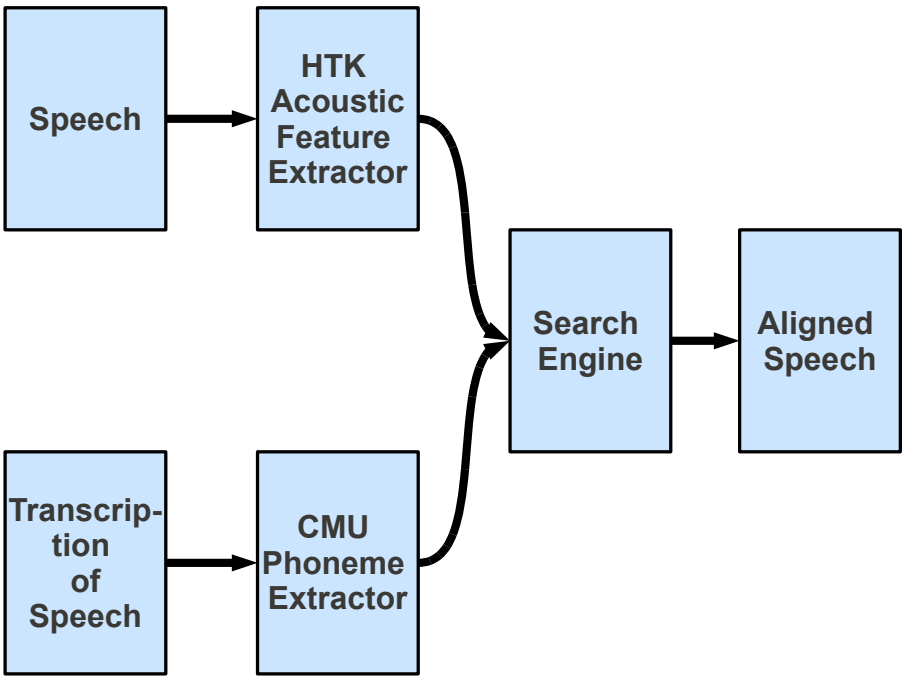


Figure 2-5: Schematic of the forced-alignment pipeline.

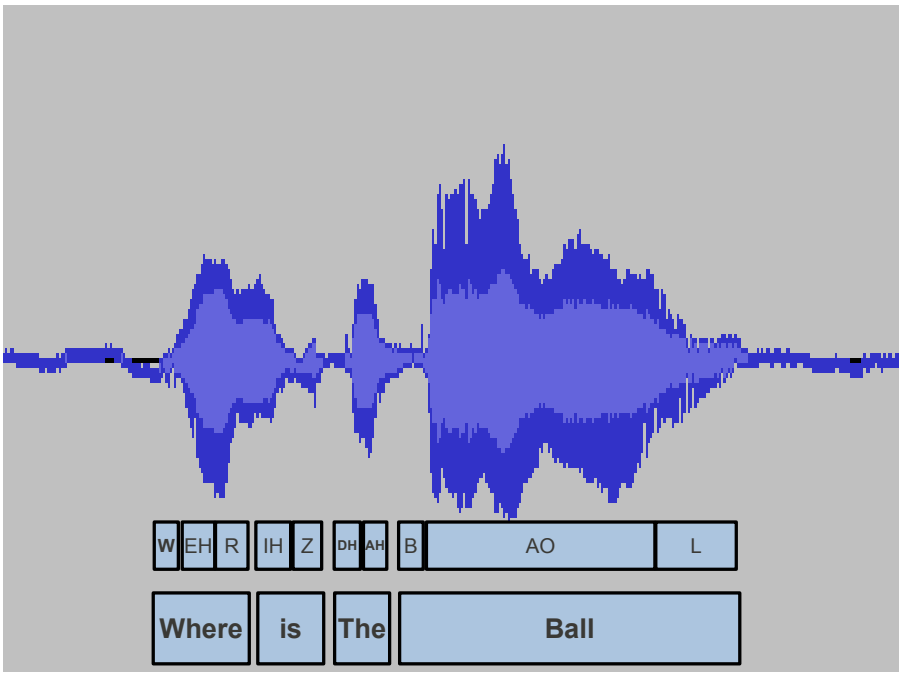


Figure 2-6: A Sample phoneme level alignment generated by the HTK forced-aligner.

2.3.5.1.1 Speaker Dependent Acoustic Models An acoustic model is a file that contains a statistical representation of each distinct phoneme that makes up a spoken word. It contains the sounds for the phonemes of each of the words used in our corpus. The Hidden Markov Model Toolkit (HTK) [38] provides tools for training acoustic models. In order to train an acoustic model, we need audio samples for each of the phonemes used in our corpus. Since there are three distinct speakers in our corpus (the three primary caregivers) we can train separate acoustic models for each speaker by training only on audio samples from their speech. These acoustic models are called speaker-dependent acoustic models since they are tuned to voice of one speaker. If trained with enough samples, speaker dependent acoustic models are better tuned to their targeted speaker and thus more accurate. Given the sheer size of our corpus, we easily had access to enough transcribed audio samples for each caregiver to train a separate acoustic model for each of the three primary caregivers in HSP.

2.3.5.1.2 Evaluation Before the forced-aligner could be used for extraction of phoneme durations we needed to evaluate its performance on the HSP dataset. In order to evaluate the aligner, we manually aligned more than 400 audio segments using the Audacity [31] software. The manual alignment was done at the word level since it is extremely difficult and time consuming for humans to do phoneme level alignment. We then measured the accuracy of our aligner using an evaluation algorithm developed by Yoshida [37]. This algorithm compares the aligned boundaries of each word from the human aligned transcripts to those of the automatically aligned transcripts. The boundaries are then classified as correct or incorrect. We defined an automatically aligned boundary as being correct if it is within some range k of the manually aligned boundary. The value of k is set to be 0.2 seconds. The accuracy of the forced-aligner is then measured by the ratio of correct alignments over all alignments.

The forced-aligner generates an acoustic score for each alignment it does. The acoustic score is a measure of how confident the system is about the the alignment. Using this score, we can automatically eliminate the worst alignments, thus increasing the aligner's overall accuracy. However, we can not just throw out words with bad alignments, we need to throw

out the whole utterance that contains the word. Figure 2-7 shows the accuracy of the aligner vs yield (measure of what percentage of utterances are used). We use the acoustic score corresponding to 90% accuracy as our cutoff for our future alignments. The 90% cutoff was selected by searching all possible cutoff values (from 1 to 100 percent yield) using our parallelized infrastructure. For each cutoff value, we performed a univariate correlation analysis to calculate the correlation between duration generated using that cutoff value and AoA. We then selected the threshold which produced the largest correlation at 85% yield. The accuracy of the aligner is more than 90% at 85% yield. Figure 2-10(c) shows the correlation between AoA and duration at each yield cutoff. The density of our dataset allows us to sacrifice about 15% of our utterances for greater accuracy.

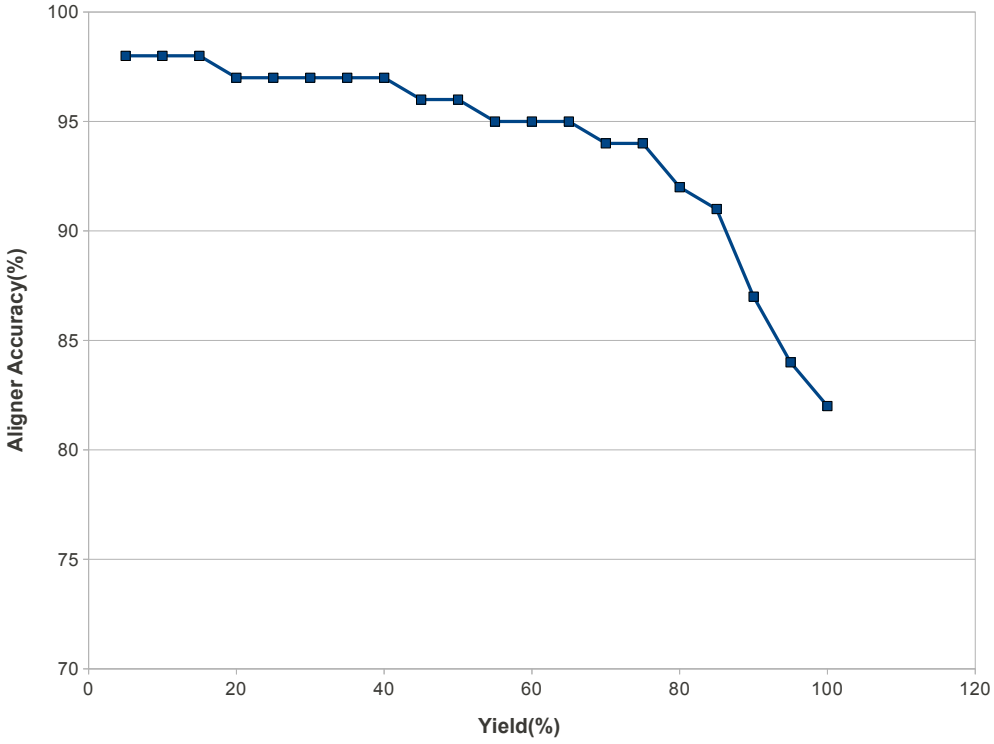


Figure 2-7: Accuracy of the aligner vs. yield. The plot shows how much data needs to be thrown out in order to achieve different levels of accuracy.

2.3.6 Fundamental Frequency (F0)

The fundamental frequency predictor is the measure of a word's change in fundamental frequency (F0) relative to the utterance in which it occurred. We first extracted the F0 contour for each utterance in the corpus using the PRAAT system [1]. Figure 2-8 shows a sample F0 contour for an utterance. We needed to come up with a formula for calculating and discretization the change in F0 from the F0 contour. We came up with a total of seven different operations for encoding the F0 predictor from the contour. Below we describe each of these operations.

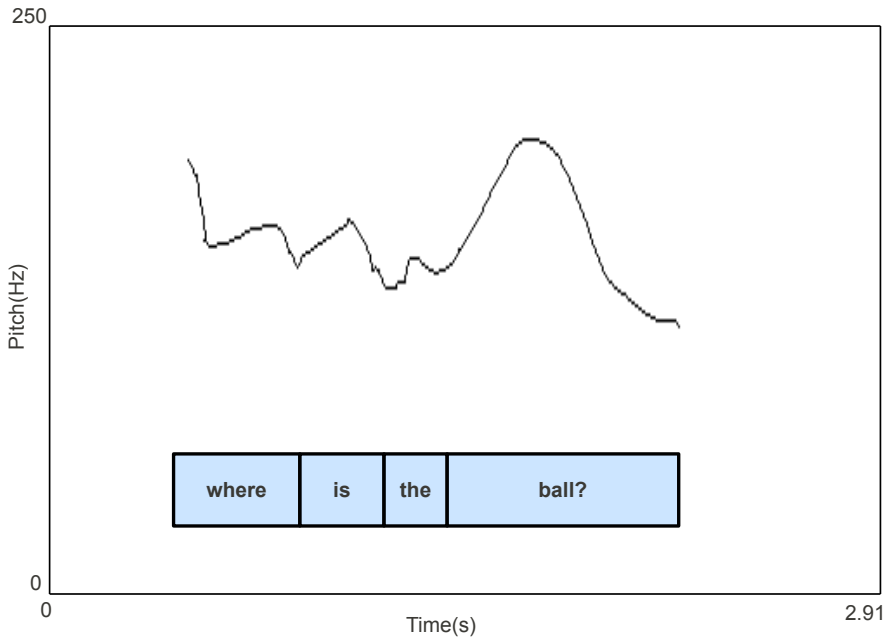


Figure 2-8: Sample F0 contour extracted from PRAAT with aligned text transcript.

UTTERANCE-MAX-CHANGE

This operation measures the change in F0 from the minimum of the F0 contour to the maximum of the F0 counter of the whole utterance(Equation (2.1)).

$$\max(\mathrm{F0}_{\mathrm{utt}}) - \min(\mathrm{F0}_{\mathrm{utt}}) \tag{2.1}$$

WORD-MAX-CHANGE

This operation measures the change in F0 from the minimum of the F0 contour of the word that we are interested in to the maximum of the F0 counter of that word(Equation (2.2)).

$$\max(\mathrm{F0}_{\mathrm{word}}) - \min(\mathrm{F0}_{\mathrm{word}}) \tag{2.2}$$

UTTERANCE-RATE-OF-CHANGE

This operation measures the absolute rate of change of F0 from the minimum of the F0 contour to the maximum of the F0 counter of the whole utterance(Equation (2.3)). This is the same as UTTERANCE-MAX-CHANGE normalized by time.

$$\frac{\max(\mathrm{F0}_{\mathrm{utt}}) - \min(\mathrm{F0}_{\mathrm{utt}})}{t_{\max(\mathrm{F0}_{\mathrm{utt}})} - t_{\min(\mathrm{F0}_{\mathrm{utt}})}} \tag{2.3}$$

WORD-RATE-OF-CHANGE

This operation measures the absolute rate of change of F0 from the minimum of the F0 contour of the word that we are interested in to the maximum of the F0 contour of that word(Equation (2.4)). This is the same as WORD-MAX-CHANGE normalized by time.

$$\frac{\max(\mathrm{F0}_{\mathrm{word}}) - \min(\mathrm{F0}_{\mathrm{word}})}{t_{\max(\mathrm{F0}_{\mathrm{word}})} - t_{\min(\mathrm{F0}_{\mathrm{word}})}} \tag{2.4}$$

UTTERANCE-VARIANCE

This operation measures the variance of the F0 contour of the whole utterance.

WORD-VARIANCE

This operation measures the variance of the F0 contour of the word in question.

UTTERANCE-WORD-CHANGE

This operation measures the absolute difference between the average F0 of the word and the average F0 of the utterance it is embedded in(Equation (2.5)).

$$\left|\overline{\text{F0}}_{\text{word}} - \overline{\text{F0}}_{\text{utt}}\right| \tag{2.5}$$

Similar to the previous predictors, we wanted the operation that generates the greatest correlation between F0 and AoA. Therefore, using our parallelized infrastructure, we searched through all the possible combinations of these seven operations(total of 5040 possible combinations). For each combination, we performed a univariate correlation analysis to calculate the correlation between F0 calculated using that operation and AoA. We then selected the combination of operations which produced the largest correlation. Figure 2-10(d) shows the sorted correlations between AoA and F0 for each different combination of operations(5040 total).

The combination of operations that produced the highest correlation is shown in Equation (2.6).

$$\alpha_0 * \left|\overline{\text{F0}}_{\text{word}} - \overline{\text{F0}}_{\text{utt}}\right| + \alpha_1 * \left|\frac{\max(\text{F0}_{\text{word}}) - \min(\text{F0}_{\text{word}})}{t_{\max(\text{F0}_{\text{word}})} - t_{\min(\text{F0}_{\text{word}})}}\right| \tag{2.6}$$

The first term in the equation captures the change in F0 for the word relative to the utterance in which it's embedded. The second term captures the maximum change in F0 within the word. α_0 and α_1 are constants which were also set by searching to maximize the correlation. Their values were set to be $\alpha_0 = 0.36$ and $\alpha_1 = 0.64$.

2.3.7 Intensity

The intensity predictor is the measure of a word's change in intensity relative to the utterance in which it occurred. We first extracted the intensity contour for each utterance in the corpus using the PRAAT system. Figure 2-9 shows a sample intensity contour for an utterance. As in with F0, we needed to come up with a formula for calculating and discretization the change in intensity from the intensity contour. We searched through all possible combinations of the same seven operations that we used for F0. Figure 2-10(e) shows the sorted correlation between AoA and intensity for each different combination of operations(5040 total). The combination of operations that produced the highest correlation was the same as the one for F0, as shown in Equation (2.7). As with F0, α_0 and α_1 were also set by searching to maximize the correlation. Their values were set to be $\alpha_0 = 0.45$ and $\alpha_1 = 0.55$.

$$\alpha_0 * \left|\overline{\text{intensity}}_{\text{word}} - \overline{\text{intensity}}_{\text{utt}}\right| + \alpha_1 * \left|\frac{\max(\text{intensity}_{\text{word}}) - \min(\text{intensity}_{\text{word}})}{t_{\max(\text{intensity}_{\text{word}})} - t_{\min(\text{intensity}_{\text{word}})}}\right| \quad (2.7)$$

2.3.8 Time-of-Day

The time-of-day predictor is different from other predictors as it does not code anything in caregiver speech. The time-of-day predictor measures the average time of day at which each word was used by the caregivers in child available speech. The child is most likely more receptive at certain times of the day and we hoped to capture this phenomenon by our time-of-day predictor variable.

To be consistent with the direction of correlation for other variables (a negative correlation with the AoA) , we transformed time of day to scale from 0 to 1 with 0 being 12:00AM and 1 being 11:59PM.

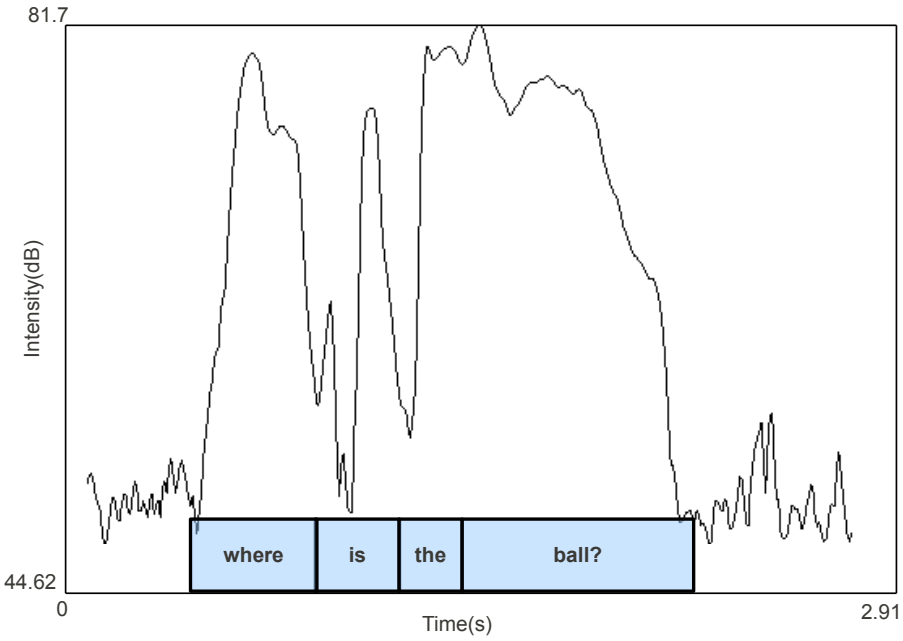


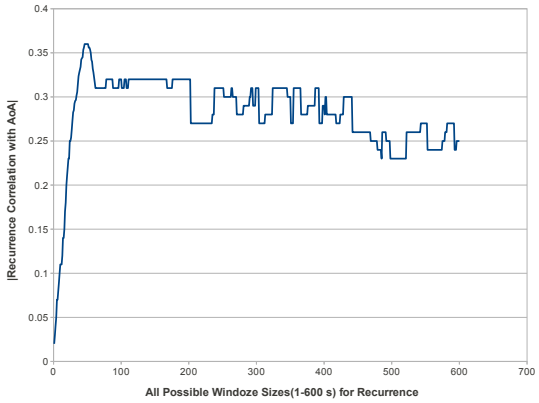
Figure 2-9: Sample intensity contour extracted from PRAAT with aligned text transcript.

2.3.9 Control Variable: Day of Week

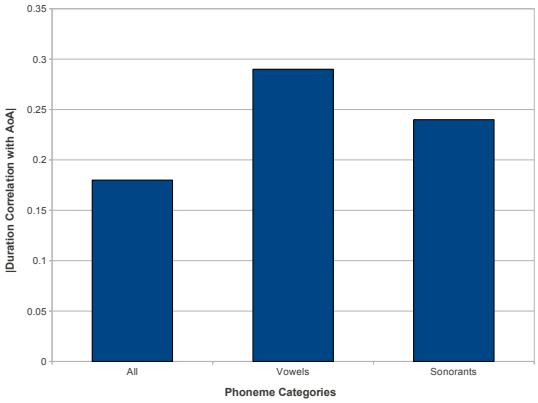
In addition to the seven predictor variables we also looked at day-of-week as a control variable. Intuitively, it seems unlikely that the day of week would have any effect on child word acquisition. So we use this variables to make sure that the correlations between AoA and our seven predictor variables are not just statistical artifacts that can be replicated with any random variable such as day-of-week.

2.4 Scripting Language for Study of HSP

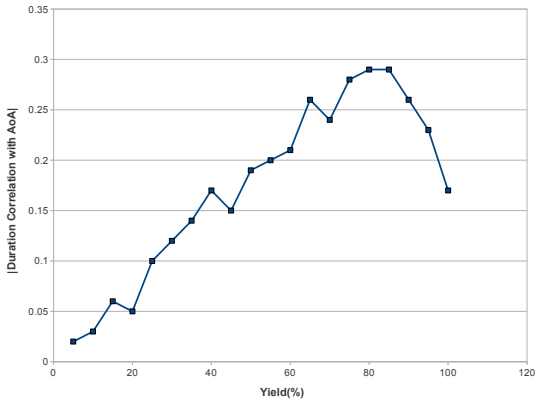
We created a scripting language that allows any user to prob many different aspects of the HSP corpus with relative ease by simply writing high level commands. This languages removes the user from all the nifty-gritty details of the processing pipelines that are needed to process and analyze the corpus. As Figure 2-11 shows, the back-end of the scripting



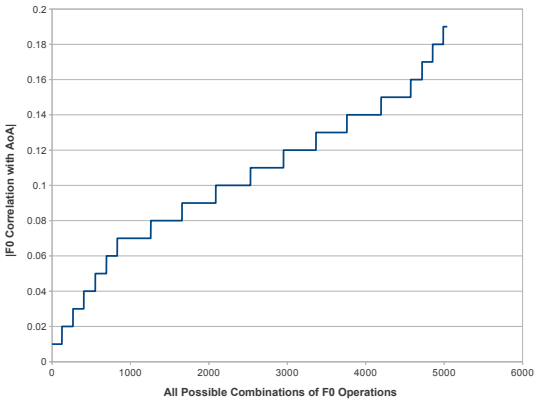
(a) Recurrence optimization



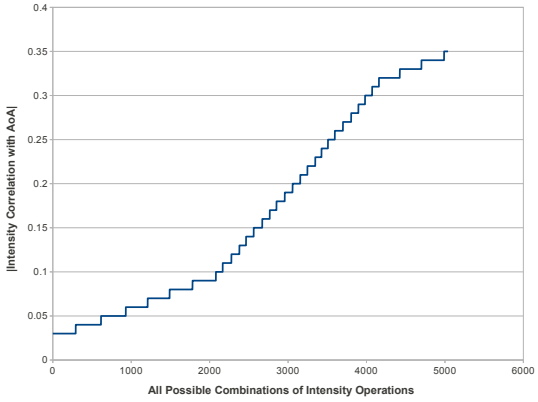
(b) Duration optimization



(c) Aligner/Duration optimization



(d) F0 optimization.



(e) Intensity optimization

Figure 2-10: Each subplot shows one of the predictor variables' optimization graph. Each subplot shows the absolute value of the correlations between AoA and a predictor variable for each of the possible operational definitions of that variable. The definition with the highest correlation was picked for each predictor variables. For clarity, some subplots show the operations sorted by the their correlations.

language is attached to the variable extractor, optimizer, parallelizer, statistical analyzer and the whole of the HSP corpus. All of that is however hidden from the user. When a program is created and run by the user, an interpreter translates the program into a sequence of commands which it then executes. The interpreter has access to all of the HSP corpus and all the tools described so far in this thesis (variable extractors, parallelizer, etc). The interpreter automatically figures out which tools to utilize in order to efficiently execute the user's program. The overall pipeline of the system can be seen in Figure 2-11.

The language consists of three different types of modules: variable modules, filter modules and processor modules. Each of these categories are described in detail below.

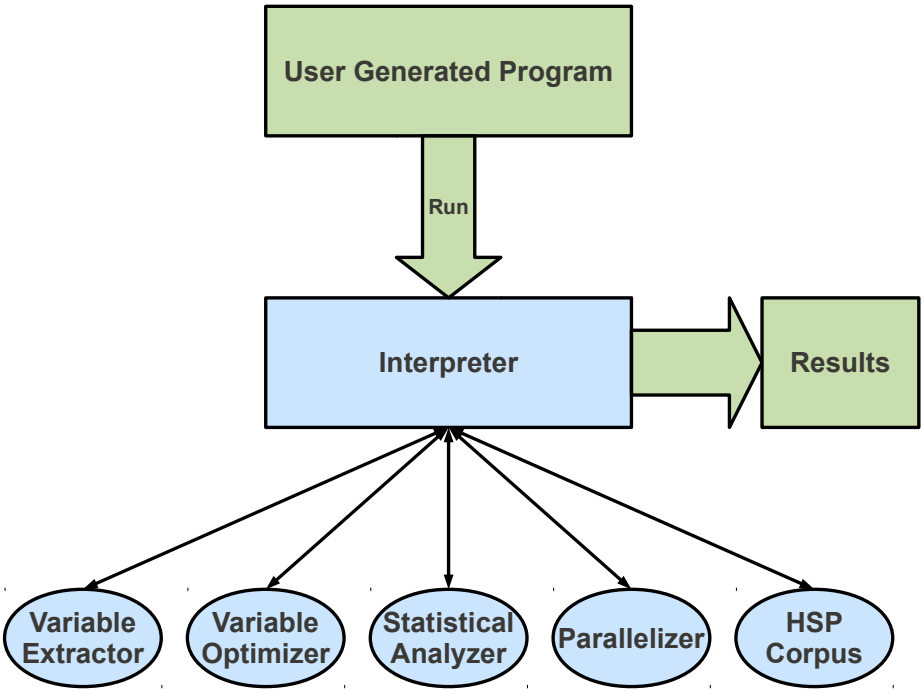


Figure 2-11: Schematic of the processing pipeline for the HSP scripting language. Only green parts of the diagram is visible to the user. The blue parts are all hidden from the user.

Variable Modules

Each variable module represents a predictor variable. Currently there are a total of 8 variable modules available to be used in our scripting language (7 predictor variable plus 1

control variable). For variables that have parameters that need to be set (such as the time window parameter in the recurrence variable) the user can either manually input values for the parameters or can opt to use the optimizer module (which is a processor module) to search for the optimal values for the parameters. The optimizer module is explained in the Processor Modules section. The description for each of the 8 variables represented by the variables modules can be found in section 3.3 of this thesis.

Filter Modules

Filter modules can be used to filter the HSP corpus to look for specific word classes, caregivers, time spans and a whole range of other things. Currently, there are a total of 5 filter modules available to be used in our scripting language. All filter modules are described below.

PART-OF-SPEECH MODULE

This module filters the corpus so that only words with the user-specified part of speech tags are considered for processing and analysis later on by the processor modules.

CDI MODULE

Similar to the part-of-speech module, this module filters the corpus so that only words that belong to the user-specified CDI(Communicative Development Inventories) categories are considered for processing and analysis later on by the processor modules.

TIME-RANGE MODULE

This module filters the corpus so that only the data that falls within the user-specified range is considered for processing and analysis later on by the processor modules.

UTTERANCE-LENGTH MODULE

This module filters the corpus so that only caregiver utterances whose length is within the user-specified range are considered for processing and analysis later on by the processor modules.

CAREGIVER MODULE

This module filters the corpus so that only utterances spoken by the user-specified caregivers are considered for processing and analysis later on by the processor modules.

Processor Modules

Processor modules are used to process and analyze the variables represented by the variable modules. Currently, there are a total of 4 processor modules available to be used in our scripting language. All processor modules are described below.

OPTIMIZE MODULE

This module, when invoked, will optimize all the unspecified parameters in the variables selected by the user through the variable modules.

CORRELATION MODULE

This module calculates the correlations for the variables specified and filtered by the user. This module has one parameter which allows the user to either generate correlations between AoA and the variables or to generate cross-correlations between the variables.

SIGNIFICANCE MODULE

This module calculates the statistical significance (p values) for variables specified and filtered by the user. Similar to the correlation module, this module has one parameter which allows the user to either generate p values for correlations between AoA and the variables or to generate p values for cross-correlations between the variables.

RAWDATA MODULE

This module generates raw data for all the user-specified and filtered variables.

The scripting language is very modular in that the user can mix and match any number of filters, variables and processors together to create a program. As mentioned, the user is not

at all involved in the processing of the program. The program is run automatically on our parallelized system and the results are returned to the user. Depending on the processor modules that the user selected, the results can be anything from a single correlation value to raw data on the F0 of all utterances in HSP.

Figure 2-12 shows a sample program created by a user. The program is run linearly from top to bottom. In the example provided, the user is asking for three variables: frequency, recurrence(with window size of a 100 seconds) and duration from 9–24 months, using CAS of all nouns in the child's vocabulary. The user then wants the parameters for all the variables that have not been manually set to be optimized and their correlations with AoA returned. The program's output will be four correlation values, one for each variable and one for the linear combination of all three variables.

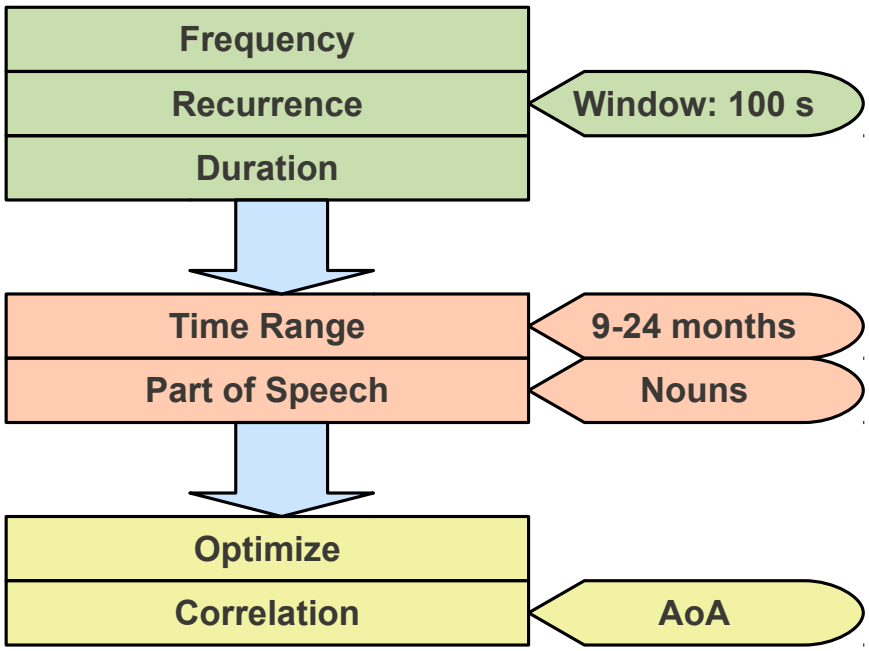


Figure 2-12: Visualization of a sample program created using the scripting language. Variable modules are in green, filter modules are in orange and processor modules are in yellow. The user is asking for three variables: frequency, recurrence(with window size of a 100 seconds) and duration from 9-24 months, using CAS of all nouns in the child's vocabulary. The user then wants the parameters for all the variables that have not been manually set to be optimized and their correlations with AoA returned. The programs output will be 4 correlation values, 1 for each variable and 1 for the combination of all three variables.

Chapter 3

Correlation Analysis

In this section we will go over the correlation of each of the seven variables we coded in CAS with AoA. We will also look at the best linear combination of these of these seven variables. Relations between input-uptake by children have been previously investigated in connection between frequencies of words in CAS and the age at which they are acquired [11, 9]. Our goal here was to replicate this type of analysis not only on frequency but also the other six variables that we coded in caregiver speech. To do this, we regressed the AoA for each word in the child's productive vocabulary against the value of each of the seven variables of that word in all the child available speech.

All correlations between AoA and the predictor variables were negative and highly significant (all p-values less than .001) though their magnitude varied. Correlations with recurrence and intensity were largest, while correlation with F0 was smallest.

3.1 Frequency

Replicating previous results in the literature [11] there was a highly significant negative correlation between frequency and AoA ($r = -.23$,$p < .001$), indicating that words that were more frequent in the child's input were acquired earlier. This correlation was mediated

by the syntactic category of the words being examined [9]. Nouns were highly correlated with AoA; verbs were considerably less so, likely due to other factors mediating acquisition [6]. Table 3.1 shows the correlations for each category in the child's speech. Figure 3-1(a) shows the scatter plot of AoA vs frequency across all caregivers and word categories.

Table 3.1: Pearson's r values measuring the correlation between age of acquisition and frequency for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Frequency	-.34**	-.48**	-.14	-.23**

3.2 Recurrence

There was also a highly significant negative correlation between recurrence and AoA ($r = -.37$,$p < .001$), indicating that words that were used by the caregivers in more dense chunks were acquired earlier. As with frequency this correlation was also mediated by the syntactic category of the words being examined. All categories were highly correlated with AoA with verbs having the highest correlation. Table 3.2 shows the correlations for each category in the child's speech. Figure 3-1(b) shows the scatter plot of AoA vs recurrence across all caregivers and word categories.

Table 3.2: Pearson's r values measuring the correlation between age of acquisition and recurrence for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Recurrence	-.41**	-.39**	-.45**	-.37**

3.3 Mean Length of Utterance(MLU)

MLU (recall we are actually looking at 1/MLU) and AoA were also significantly negatively correlated ($r = -.25$,$p < .001$). This indicates that words that were used in less complex

utterances by caregivers were acquired earlier by the child. Adjectives were highly correlated with AoA; while nouns and verbs were considerably less so. Table 3.3 shows the correlations for each category in the child's speech. Figure 3-1(c) shows the scatter plot of AoA vs MLU across all caregivers and word categories.

Table 3.3: Pearson's r values measuring the correlation between age of acquisition and 1/MLU for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
1/MLU	-.30**	-.14*	-.17*	-.25**

3.4 Duration

Duration and AoA were significantly negatively correlated ($r = -.29$,$p < .001$), indicating that words that were often spoken with relatively greater emphasis (through elongated vowels) were acquired earlier. Adjectives were highly correlated with AoA; while nouns and verbs were considerably less so. Table 3.4 shows the correlations for each category in the child's speech. Figure 3-1(d) shows the scatter plot of AoA vs duration across all caregivers and word categories.

Table 3.4: Pearson's r values measuring the correlation between age of acquisition and duration for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Duration	-.44**	-.13*	-.19*	-.29**

3.5 Fundamental Frequency(F0)

Though weaker than the rest of the variables, F0 and AoA were also significantly negatively correlated ($r = -.19$,$p < .001$), again indicating that words that were often spoken with relatively greater emphasis (through change in F0) were acquired earlier. Similar to duration, adjectives were highly correlated with AoA; while nouns and verbs were considerably less

so. Table 3.5 shows the correlations for each category in the child's speech. Figure 3-1(e) shows the scatter plot of AoA vs F0 across all caregivers and word categories.

Table 3.5: Pearson's r values measuring the correlation between age of acquisition and F0 for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
F0	-.27*	-.17*	-.09	-.19**

3.6 Intensity

From all the prosodic variables (duration, f0 and intensity), intensity had the strongest negative correlation with AoA ($r = -.35$,$p < .001$), once again indicating that words that were often spoken with relatively greater emphasis (through change in intensity) were acquired earlier. Similar to the other prosodic variables, adjectives were highly correlated with AoA; while verbs were considerably less so. Table 3.6 shows the correlations for each category in the child's speech. Figure 3-1(f) shows the scatter plot of AoA vs intensity across all caregivers and word categories.

Table 3.6: Pearson's r values measuring the correlation between age of acquisition and intensity for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Intensity	-.43**	-.37**	-.20*	-.35**

3.7 Time-of-Day

Time-of-day, the only predictor variable that did not code anything in caregiver speech, was also significantly correlated with AoA ($r = -.21$,$p < .001$). This could possibly indicate that the words that were used on average at certain times of the day were acquired earlier by the child, which could mean that the child is more receptive to word learning at certain times of the day. However, as we will show later on in this thesis, when all the variables are combined

in a linear combination, the time-of-day variable becomes statistically insignificant. Table 3.7 shows the correlations for each category in the child's speech. Figure 3-1(g) shows the scatter plot of AoA vs time-of-day across all caregivers and word categories.

Table 3.7: Pearson's r values measuring the correlation between age of acquisition and time-of-day for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Time-of-Day	-.30*	-.16*	-.32**	-.21**

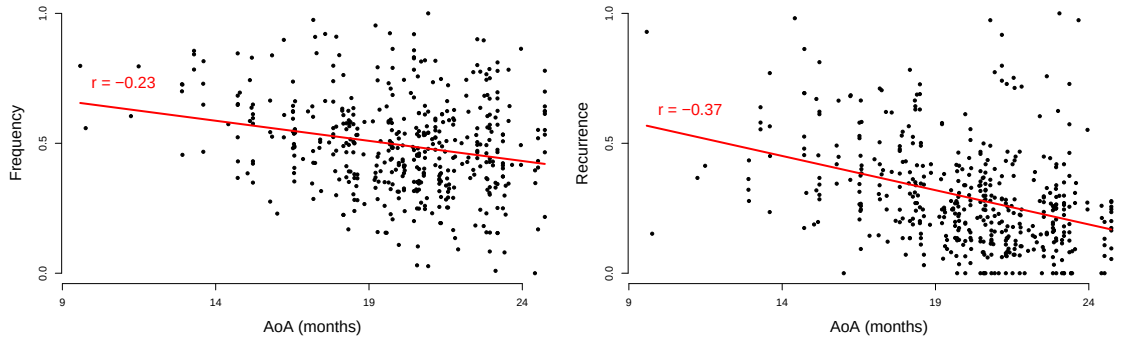
3.8 Day-of-Week

Finally, the correlation between the control variable day-of-week and AoA was neither strong nor statistically significant ($r = .04$,$p = .29$), indicating that the strong, significant correlations between AoA and our seven predictor variables are not just statistical artifacts.

3.9 Summary of Univariate Correlation Analysis

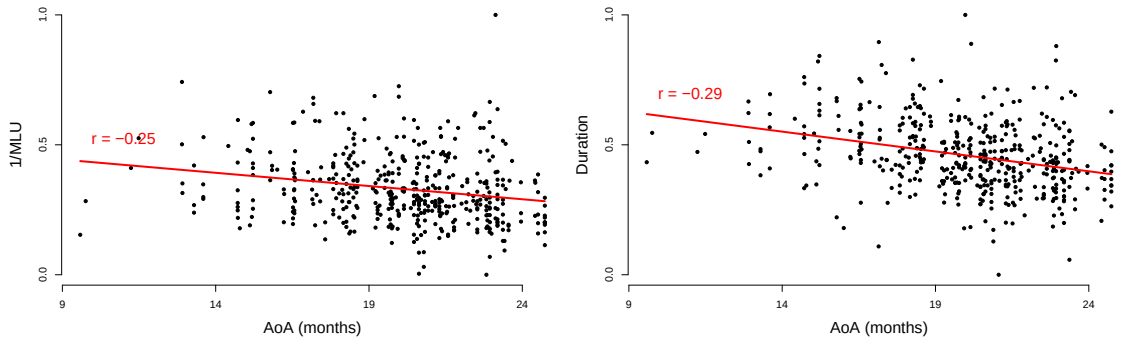
Table 3.8 shows the correlations for each of the seven predictor variables for each category in the child's speech. It is interesting to note that all the prosodic variables (duration, f0 and intensity) were highly correlated with adjectives while frequency of word use and recurrence were highly correlated with nouns and verbs respectively. The strong correlation between the prosodic variables and AoA of adjectives suggests that emphasis on adjectives has a great impact on the acquisition of those adjectives by the child. It is also interesting to note that each of the word categories (nouns, verbs and adjectives) have the highest correlation with at least one of predictive variables, possibly indicating that the child uses different signals in caregiver speech for learning different word categories.

As mentioned, the correlations obtained for frequency replicate previous results in the literature [11]. Moreover, the significant correlations between AoA and the three prosodic variables(duration, fundamental frequency and intensity) agree with previous work that



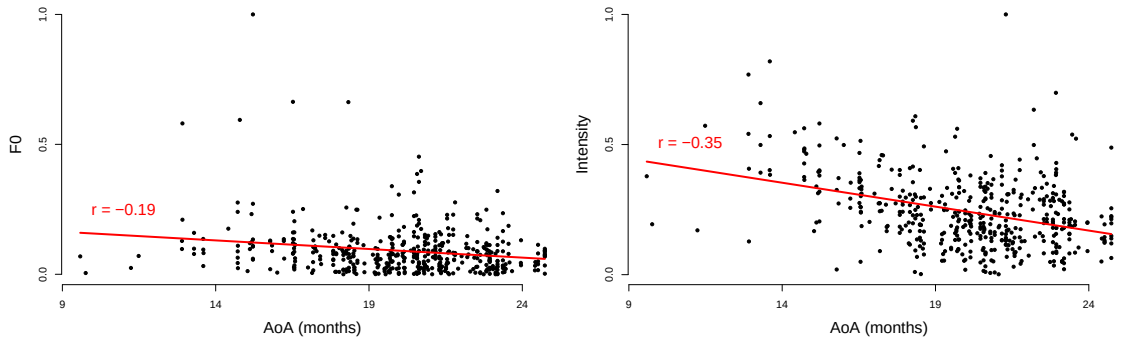
(a) Frequency

(b) Recurrence



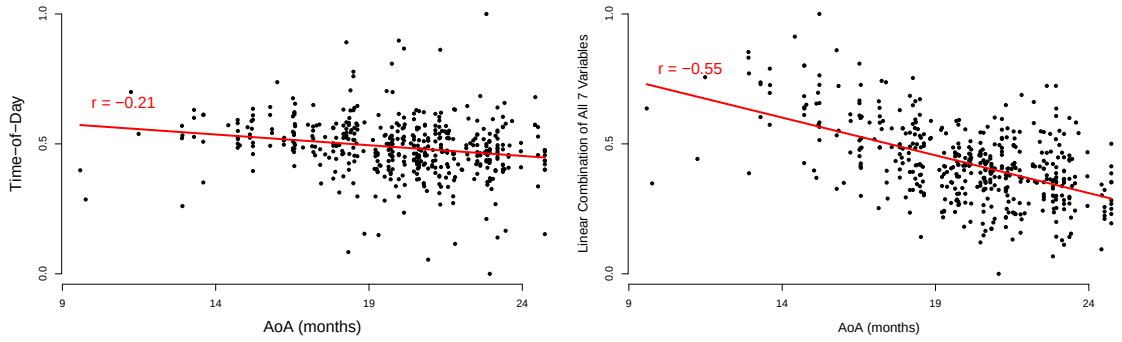
(c) MLU

(d) Duration



(e) F0

(f) Intensity



(g) Time-of-Day

(h) Best Linear Combination of All 7 Variables

Figure 3-1: Each subplot shows the univariate correlation between AoA and a particular predictor. Each point is a single word, while lines show best linear fit.

show prosodic cues in CAS contribute to acquisition of language [13, 19]. Additionally, the prominence of intensity over the other prosodic variables in our results agree with a previous study[14].

Table 3.8: Pearson's r values measuring the correlation between age of acquisition and each of the seven predictor variables for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Frequency	-.34**	-.48**	-.14	-.23**
Recurrence	-.41**	-.39**	-.45**	-.37**
1/MLU	-.30**	-.14*	-.17*	-.25**
Duration	-.44**	-.13*	-.19*	-.29**
F0	-.27*	-.17*	-.09	-.19**
Intensity	-.43**	-.37**	-.20*	-.35**
Time-of-Day	-.30*	-.16*	-.32**	-.21**

3.10 Linear Combination of All Seven Predictor Variables

We next looked at the correlation between AoA and the best linear combination of all the 7 predictor variables, shown in Figure 3-1(h). The correlation between AoA and the best linear combination of the seven predictor variables was $r = -.55$ ($p < .001$). We also looked the correlation between AoA and the linear combination of the best 2, 3, 4, 5 and 6 predictor variables (best is defined as one with the highest correlation). This was done to observe how much each additional variable improves the correlation with AoA. Table 3.9 shows those correlations for each category in the child's speech.

Table 3.10 shows the statistical significance of each of the variables when linearly combined as presented in Table 3.9. As shown in Table 3.10, when linearly combined, all the variables except for time-of-day remained significant. The statistically insignificance of time of day when combined with other variables suggests that the time-of-day variable is statistically explained by our other six variables. For example, it could be that the caregivers interact more with the child at certain time of day, not that the child is more receptive to learning at those times. Since time-of-day is no longer statistically significant we do not include it

Table 3.9: Pearson's r values measuring the correlation between age of acquisition and the linear combinations of the best 2, 3, 4, 5, 6 and 7 predictor variables. Significant codes: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adj.	Nouns	Verbs	All
Recur. + Int.	-.52**	-.49**	-.47**	-.48**
Recur. + Int. + 1/MLU	-.64**	-.51**	-.47**	-.51**
Recur. + Int. + 1/MLU + F0	-.66**	-.52**	-.46**	-.53**
Recur. + Int. + 1/MLU + F0 + Dur.	-.71**	-.51**	-.46**	-.55**
Recur. + Int. + 1/MLU + F0 + Dur. + Freq.	-.72**	-.58**	-.44**	-.55**
Recur. + Int. + 1/MLU + F0 + Dur. + Freq. + ToD	-.72**	-.58**	-.45**	-.55**

in our analysis in the rest of this thesis.

The linear model shown above has two limitations. First, we found that there was significant variation in the effects of the six predictors depending on what POS a word belonged to. Second, we did not include any interaction terms. We followed up in two ways. First, in order to investigate differences in predictor values between word classes we built separate linear models for each POS. Second, we used stepwise regression to investigate interactions in our larger model.

Table 3.11 shows coefficient estimates for five linear models, each one for a different group of words. None (including the "all" model) include a predictor for POS. Coefficient estimates varied considerably across models, suggesting that different factors are most important for the acquisition of different kinds of words. For example, frequency, intensity, and inverse MLU were most important for nouns, suggesting that hearing a noun often in short sentences where it is prosodically stressed leads to earlier acquisition. In contrast, adjective AoA was best predicted by intensity, duration, and inverse MLU, congruent with reports that children make use of prosodic cues in identifying and learning adjectives [32]. Finally, both verbs and closed-class words were best predicted by recurrence, supporting the idea that the meanings of these words may be difficult to decode from context; hence frequent repetition within a particular context would be likely to help [6].

3.11 Cross Correlation Between Predictor Variables

Correlations between the predictor values are shown in Table 3.12. The largest correlations were between frequency and recurrence, frequency and intensity, and inverse MLU and duration. The correlation between frequency and recurrence is easily interpreted: the more times a word appears, the more likely it is to recur within a small window. Moreover, the correlation between MLU and duration can be explained, since in general words are shorter when the sentences are longer (similar to syllables being shorter when the words are longer). Finally, the correlation between recurrence and duration can also be interpreted: generally, words are shorter when they are repeated, as shown in a study by Fowler [5]. On the other hand, correlations between other variables like frequency and intensity are less clear.

Table 3.10: Statistical significant of each of the 7 predictor variables for linear combinations of best 2, 3, 4, 5, 6 and 7 predictor variables. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Recur.	Int.	1/MLU	F0	Dur.	Freq.	ToD
Best 2	**	**					
Best 3	**	**	**				
Best 4	**	**	**	*			
Best 5	**	**	**	*	*		
Best 6	**	**	**	*	*	*	
Best 7	**	**	**	*	*	*	not significant

Table 3.11: Coefficient estimates for linear models including data from adjectives, nouns, closed-class words, verbs, and all data. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Closed	Nouns	Verbs	All
Intercept	27.66**	25.03**	25.00**	25.93**	25.57**
Frequency	0.38	6.73	-5.84**	-0.89	-1.53*
Recurrence	-2.36	-12.02*	-1.53′	-7.47**	-2.85**
Duration	-5.22*	1.81	0.09	-2.74	-2.66*
F0	-7.43	-6.42	-2.28′	0.54	-3.42*
Intensity	-8.60*	-12.16	-4.66**	-1.56	-4.78**
1/MLU	-5.70*	-9.37	-3.71*	-5.26	-3.89**

Table 3.12: Correlation coefficients (Pearson's r) between all predictor variables. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Recurrence	Duration	F0	Intensity	1/MLU
Frequency	.36**	-.05	.19**	.35**	-.22**
Recurrence		.25**	.20**	.22**	.10*
Duration			.12*	.22**	.33**
F0				.10*	-.15*
Intensity					.02

Chapter 4

Predictive Model 1

In this chapter we look at the predictive power of each of the six variables coded in caregiver speech. If certain aspects of the child's input(i.e. any of the six variables) are shown to be predictive of his lexical development, this could help us better understand the nature of the child's word acquisition mechanisms.

To that end, we constructed a regression model which attempted to predict AoA as a function of a linear combination of predictor values. The part of speech (POS) was included as an additional predictor. We created POS tags by first identifying the MacArthur-Bates Communicative Development Inventory category [4] for each word that appeared in the CDI and generalizing these labels to words that did not appear in the CDI lists. To avoid sparsity, we next consolidated these categories into five broad POS categories: adjectives, nouns, verbs, closed-class words, and other. The inclusion of POS as a predictor significantly increased model fit ($p < .001$).

Coefficient estimates for each predictor are shown in Figure 4-1. All predictors were significant at the level of $p < .05$. The full model had $r = -0.66$, suggesting that it captured a substantial amount of variance in age of acquisition.

The largest coefficients in the model were for intensity and inverse MLU. For example, there was a four-month predicted difference between the words with the lowest inverse

MLU ("actual," "rake," "pot," and "office") and the words with the highest inverse MLU ("hi," "silver," and "hmm"). Effects of POS were significant and easily interpretable. We used nouns as the base contrast level; thus, coefficients can be interpreted as extra months of predicted time prior to acquiring a word of a non-noun POS. Closed-class words and verbs were predicted to take almost two months longer to acquire on average, while adjectives and other words were predicted to take on average less than a month longer.

Part of the work described in this section is from the paper Contributions of Prosodic and Distributional Features of Caregivers Speech in Early Word Learning which was published in the proceedings of the 32nd Annual cognitive Science Conference [34].

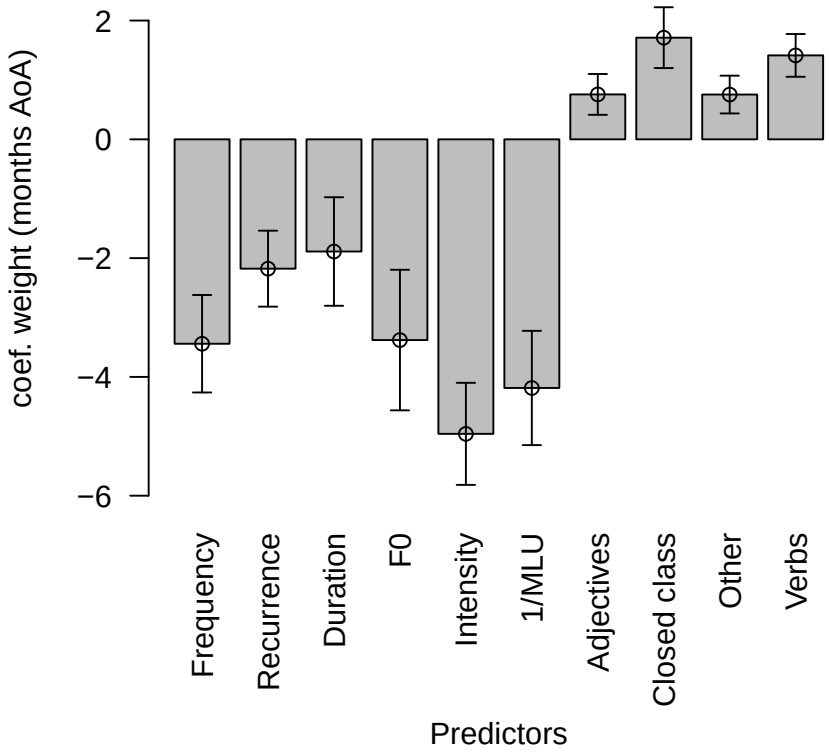


Figure 4-1: Coefficient estimates for the full linear model including all six predictors (and part of speech as a separate categorical predictor). Nouns are taken as the base level for part of speech and thus no coefficient is fit for them. Error bars show coefficient standard errors. For reasons of scale, intercept is not shown.

4.1 Evaluation of the Predictive Model 1

In order to evaluate the fitness of our predictive model we look at its predictive power through standard k-fold cross validation (K=461). This is done by ignoring one of the 461 words in the child's lexicon and redoing our entire analysis -from variable optimization to the regression analysis- on the remaining 460 words. This way the new model is completely blind to the the word that was left out. We then try to predict the age of acquisition of that particular word with our new model. Figure 4-2 shows how this is done. First, we extract and compute the six predictor variables for all CAS utterances that contain the word whose age of acquisition we are trying to predict. Note that we extract these six variables for utterances spanning the whole 9–24 timespan (as opposed to up to AoA) since the predictor is supposed to be blind to the AoA of the word whose AoA the model is trying to predict. We then use the values of the extracted variables as inputs for our predictive model which then in return predicts the age at which the word in question will be acquired. We do this for all the 461 words in our corpus. We can then compare the true age of acquisition of the 461 words with the predicted age of acquisition.

Figure 4-3 shows the relation between predicted age of acquisition (via the full predictive model including part of speech) and the true age of acquisition of words by the child. On average our new model can correctly predict the age of acquisition of a word by the child within 55 days.

4.2 Outliers

One useful aspect of plotting the data like we have in Figure 4-3 is that it makes clear which words were outliers in our model (words whose predicted age of acquisition is very different than their actual age of acquisition). Identifying outliers can help us understand other factors involved in age of acquisition.

For example, words like "dad" and "nannyname" (proper names have been replaced for privacy reasons) are learned far earlier than predicted by the model (above the line of best

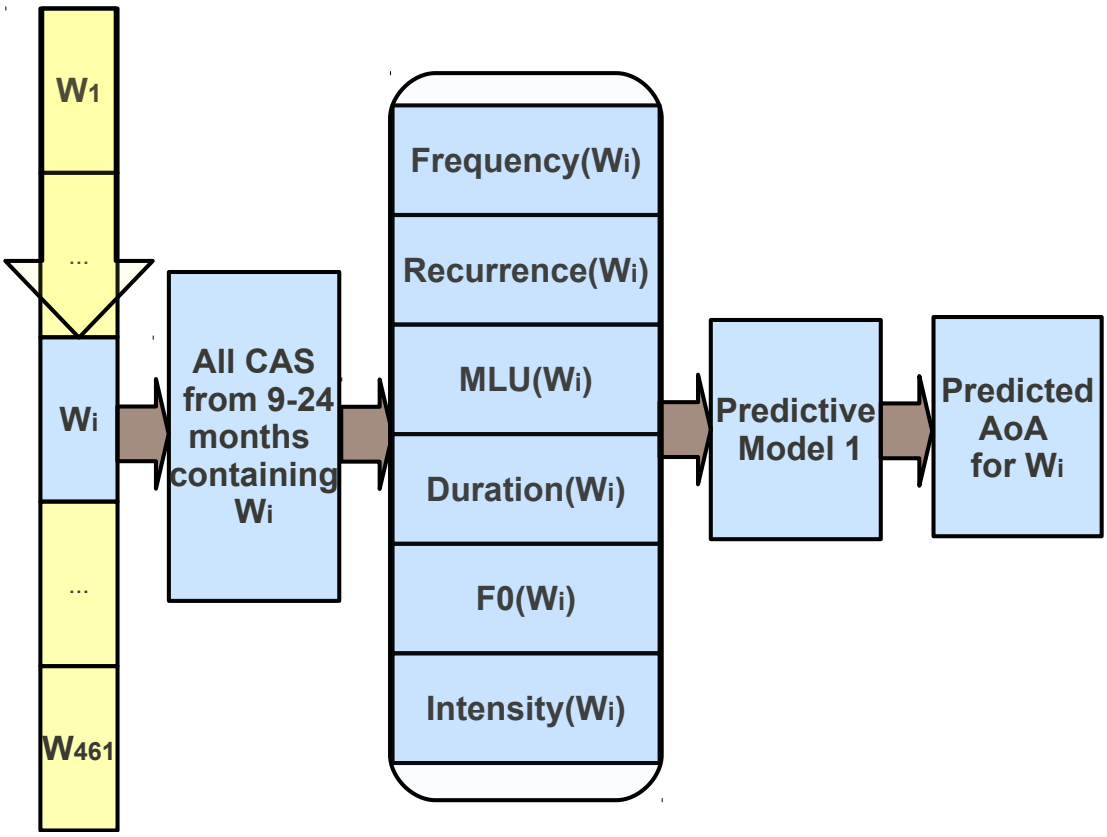


Figure 4-2: Schematic of the pipeline used for the 461-fold cross validation of predictive model 1.

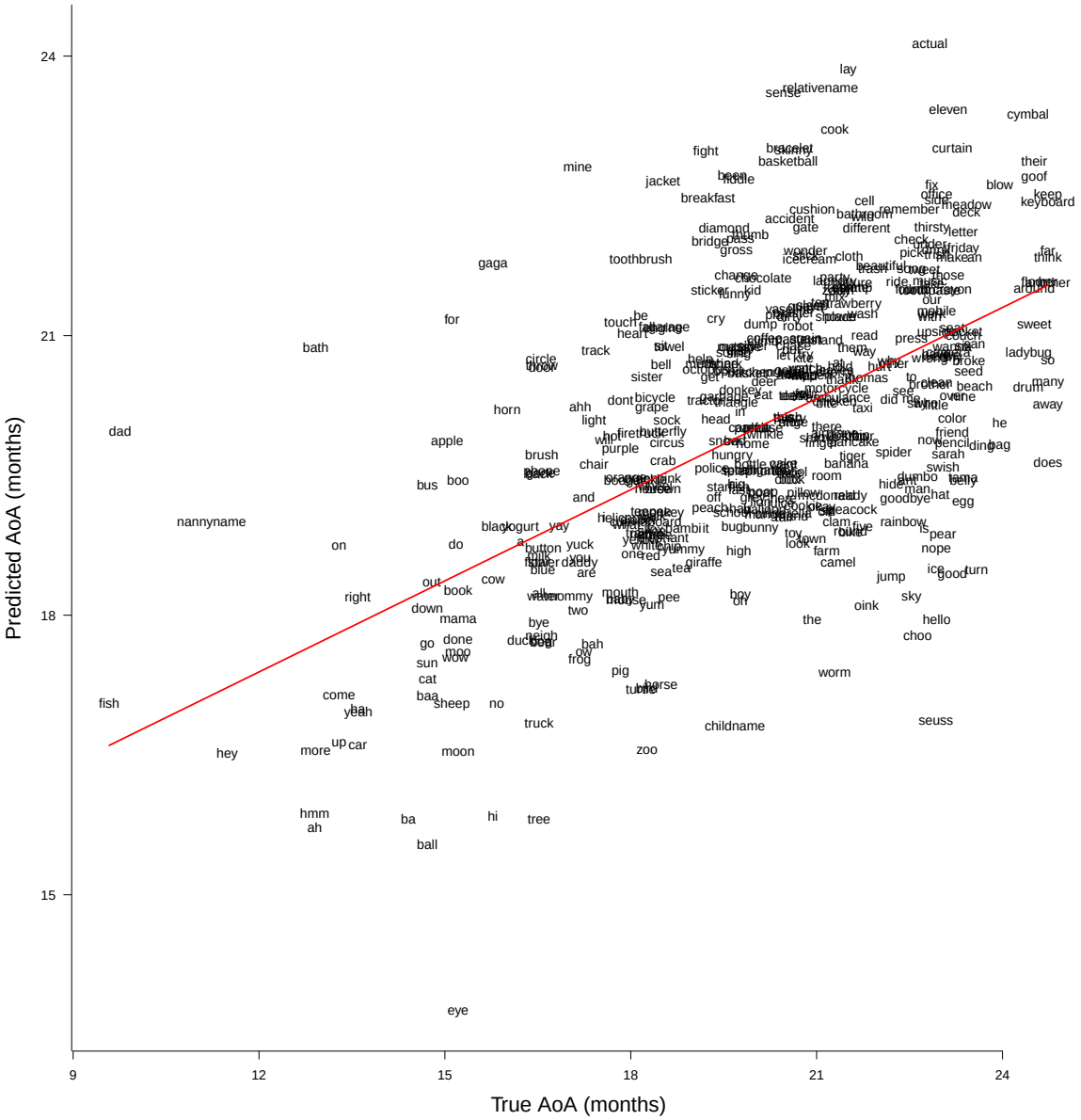


Figure 4-3: Predicted AoA by model 1 vs. true AoA

fit), due to their social salience. Simple and concrete nouns like "apple" and "bus" are also learned earlier than predicted, perhaps due to the ease of individuating them from the environment. In contrast, the child's own name is spoken later than predicted (20 months as opposed to 18), presumably not because it is not known but because children say their own name far less than their parents do.

In the future work section of this thesis will talk about how to use these errors of prediction as a starting point for understanding contextual factors influencing word learning.

Chapter 5

Limitations of Model 1

There are a few limitations to the model and analysis described in the previous section.

- First, the analysis so far assume a linear input-output model between the child and the caregivers: the caregivers provide input to the child, who then learns words. In other words, our current model treats the child as the only agent whose behavior can change. Beyond a first approximation, however, this assumption is inconsistent with other findings [27].
- Second, our current model only takes into account variables in caregiver speech, omitting the visual and social context of word learning.
- Third, our analysis uses AoA as a crude estimate for AoC.
- Fourth, our analysis use CAS as a substitute for CDS.
- Fifth, though the majority of our transcripts are fairly accurate, there is still a substantial amount of inaccurate transcripts that end up being used in our analysis.
- Sixth, the HSP corpus represents data for only a single child (N=1).

We addresses some of these issues in the remaining parts of this thesis and leave the rest for future work.

Chapter 6

Mutual Influence Between Caregivers and Child

In this section, we investigate the mutual influences between the three caregivers and the child and try to come up with a measure for the degree of adaption in this dynamic social system.

Previous work by Roy et al. [27] has shown evidence of fine lexical tuning by the caregivers. They showed that caregivers adjust the complexity of the contexts in which they present individual words (measured by looking at the MLU of the caregivers) depending on the child's understanding of those words. Using their methodology, we looked for evidence of tuning in all of the six variables that we coded in caregiver speech.

To carry out this analysis for a particular variable (e.g. fundamental frequency) for each word in the child's productive vocabulary, we extracted the variable for each month for the CAS containing that word. This resulted in a time-series for each caregiver for each word (for that particular variable). We then time-aligned these time-series so that they were aligned by age of acquisition and averaged across the words and normalized the results. Finally, the resulting curves were smoothed using locally weighted scatterplot smoothing.

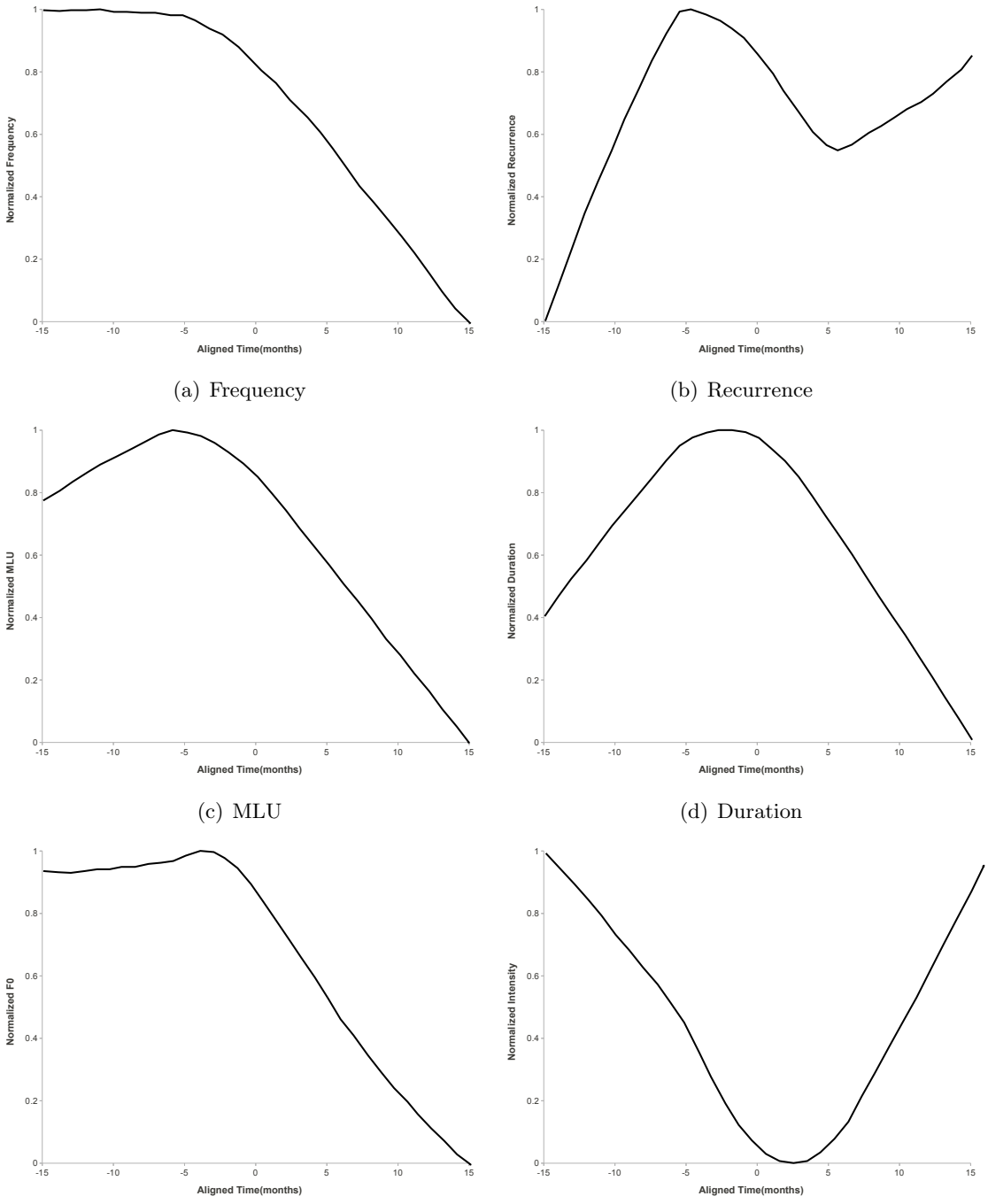
This analysis allow us to look at whether there is a consistent change in caregiver behavior

for each of the six variables before and after the AoA. It is interesting to note that this type of analysis of CAS time-series data at level of single words were not possible before the Human Speechome corpus.

Figure 6-1 shows the results of this analysis for each variable. The x-axis on the graphs is time, ranged from 15 months before AoA to 15 months after AoA. This is done because in the 9-24 months range, words with AoAs on month 9 have 15 months ahead of the AoA while words with AoAs on month 24 have 15 months before the AoA. From the graphs we can see significant evidence of caregiver "tuning" for all six variables. These six variables in caregiver speech all show significant temporal relationships with the child's lexical development, suggesting that caregivers "tune" their prosodic and distributional characteristics of their speech to the linguistic ability of the child. This tuning behavior involves the caregivers progressively shortening their utterance lengths, becoming more redundant and exaggerating prosody more when uttering particular words as the child gets closer to the AoA of those words and reversing this trend as the child moves beyond the AoA. The tuning behavior is remarkably consistent across caregivers and variables, all following a very similar pattern.

6.1 Measuring Degree of Adaption

We next tried to develop a metric for measuring degree of adaption between the caregivers and the child shown in Figure 6-1. To do this, we first came up with an upper and lower bound for each of the graphs. We came up with these bounds by swapping the AoAs of all the 461 words randomly and then regenerating the curves. We went over billions of possible combinations by utilizing our parallelized infrastructure. For each permutation we measured the "tuning score" of the generated graph. We developed the tuning score as a crude estimate of the how well the graph shows tuning behavior by the caregivers. The formula for calculating the tuning score is shown in Equation (6.1). As shown in Figure 6-2, the score is calculated by first measuring the slopes of the line between the start point of the graph and the maximum and the line between the maximum and the end point of



(a) Frequency

(b) Recurrence

(c) MLU

(d) Duration

(e) F0

(f) Intensity

Figure 6-1: The mutual influence curves for each of the 6 predictor variables.

the graph. These two slops are then added to generate the tuning score. Though crude, this score is a good measure of the change in the caregivers' behavior over time.

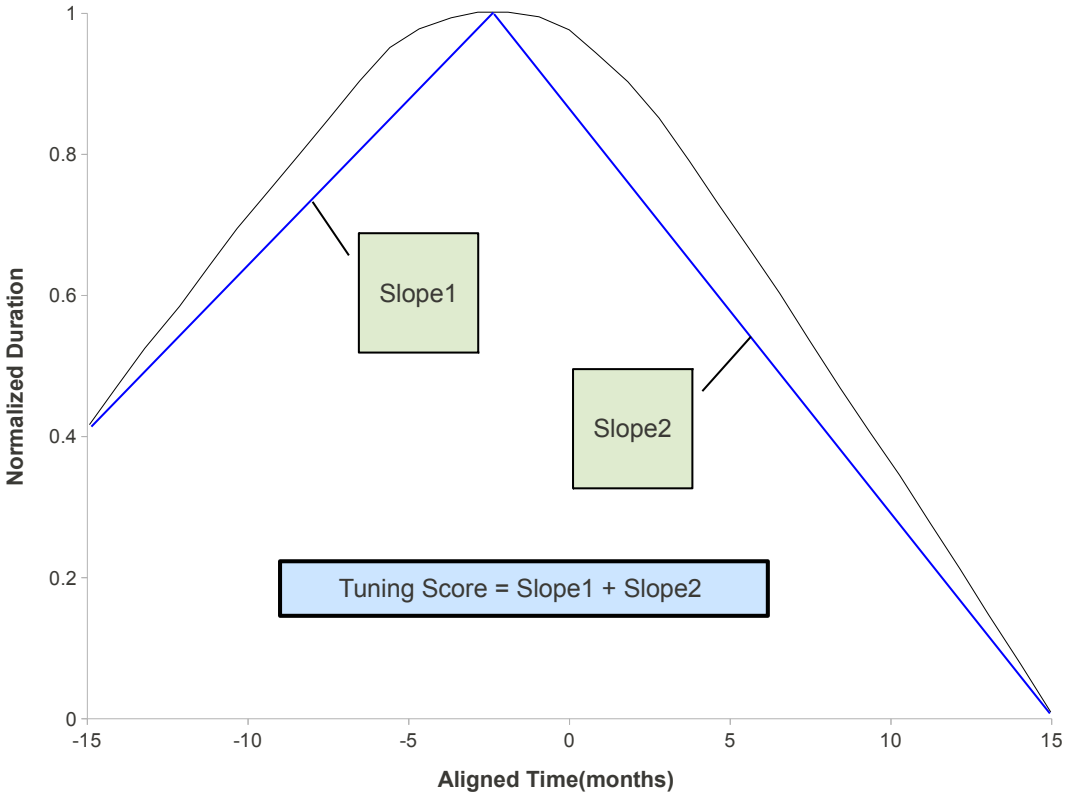
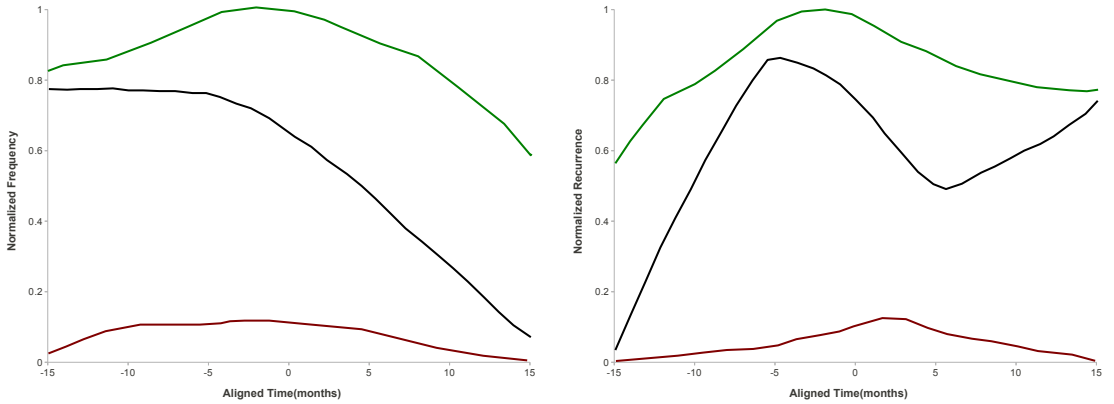


Figure 6-2: An example of the method used for calculating the tuning scores of mutual influence curves. Slopes 1 and 2 are used in Equation (6.1) which calculates the tuning score.

$$\text{TuningScore} = \left|\frac{y_{max} - y_{start}}{x_{max} - x_{start}}\right| + \left|\frac{y_{max} - y_{end}}{x_{max} - x_{end}}\right| \tag{6.1}$$

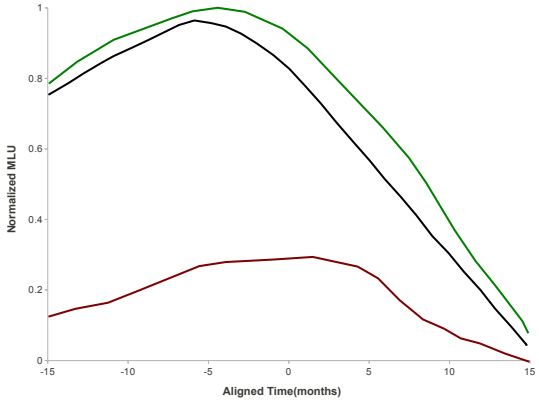
The tuning scores of the curves were then used to pick the upper and lower bound curves for each of the graphs in Figure 6-1. The curve with the highest tuning score was selected as the upper bound and the curve with the lowest score was selected as the lower bound. Figure 6-3 shows each of the curves with their corresponding bounds.

We then used these bounds to come up with a metric for measuring the degree of adaption

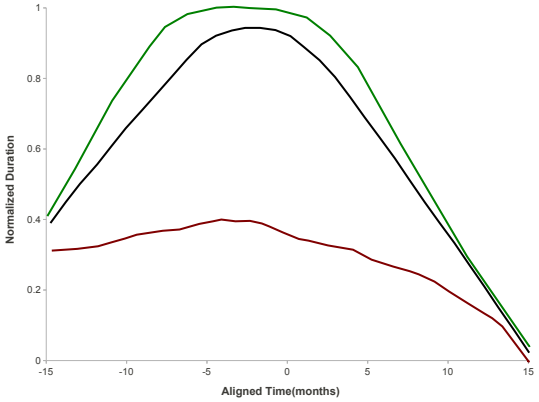


(a) Frequency

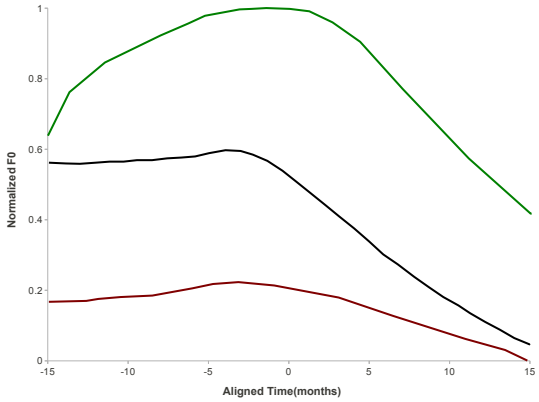
(b) Recurrence



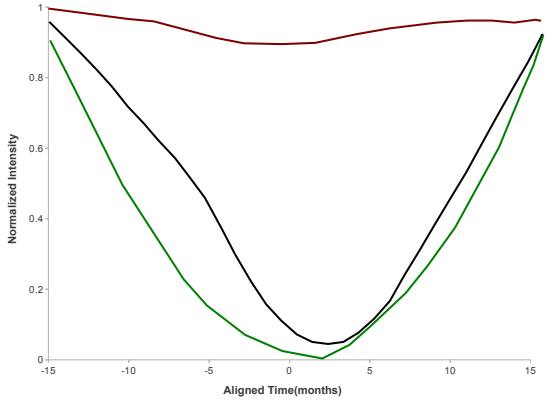
(c) MLU



(d) Duration



(e) F0



(f) Intensity

Figure 6-3: Mutual influence curves and their upper and lower bounds. The original curves are in black, the upper bounds are in green and the lower bounds are in red.

between caregivers and the child. Equation (6.2) shows the metric that was developed. Figure 6-4 is an example of this metric being applied. In Equation (6.2), A1 is the area between the upper bound and the original curve(green area in Figure 6-4) while A2 is the area between the original curve and the lower bound(red area in the figure).

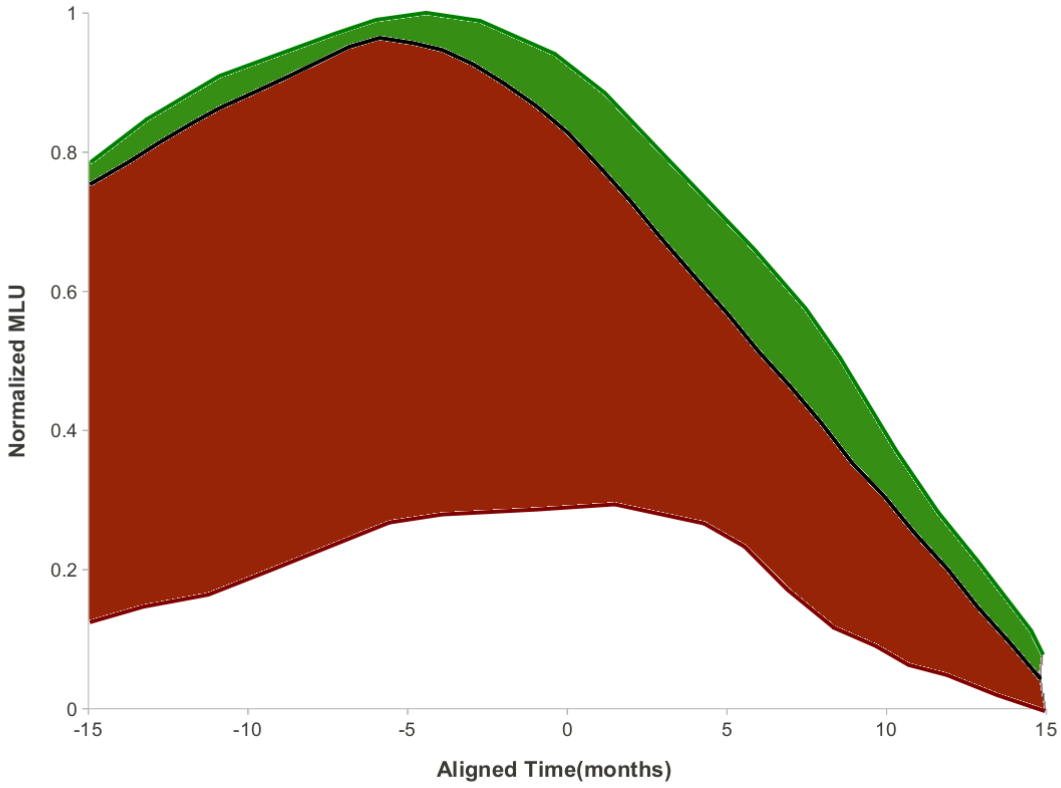


Figure 6-4: An example of how the adaption score is calculated from the mutual influence curves. The green region is the area between the mutual influence curve and its upper bound while the red region is the area between the mutual influence curve and its lower bound. The adaption score is then calculated using Equation (6.2)

$$\text{AdaptionScore} = 1 - \frac{\text{A1}}{(\text{A1} + \text{A2})} \tag{6.2}$$

We call this metric the "Adaption Score". The adaption score will be between 0 and 1 for all curves. A score 1 means perfect adaption while a score of 0 means no adaption. The adaption score would be 1 only if A1- the area between the upper bound and the original

curve- was zero, meaning that the original curve was the same as the upper bound, which means it had the maximum possible tuning score. On the other hand, the adaption score would be 0 only if A2- the area between the original curve and the lower bound- was zero, meaning that the original curve was the same as the lower bound, which means it had the minimum possible tuning score. Table 6.1 shows the adaption score for all caregivers and predictor variables.

Table 6.1: Adaption score of each of the caregivers for all the predictor variables.

	Frequency	Recurrence	1/MLU	Duration	F0	Intensity
Caregiver 1	.57	.70	.86	.66	.60	.81
Caregiver 2	.61	.60	.87	.72	.63	.77
Caregiver 3	.70	.72	.92	.69	.50	.72
All	.68	.73	.89	.77	.60	.80

6.2 Effects of Caregiver Adaption on Predictive Power of Variables

We looked at the whether the adaption scores for the variables were at all related to the correlation between the variables and AoA. Table 6.2 shows these measure side by side for each variable. In general variables with the highest degree of caregiver adaption were also more strongly correlated with the AoA; this was quantified by calculating the correlation between variable adaption scores and variable correlations with AoA for each variable. The correlation overall(all caregivers combined) was highly significant($r = -0.42$, $p < 0.1$). In other words, the predictive power of our model was highest for variables with high degree of caregiver adaption. This means that stronger tuning between the caregivers and the child for a particular variable makes that variable more predictive of the AoA of words by the child. This might be evidence that the child utilizes variables that are better tuned by the caregiver more than ones that are not as highly tuned. Please note that this analysis is preliminary and needs to be further studied.

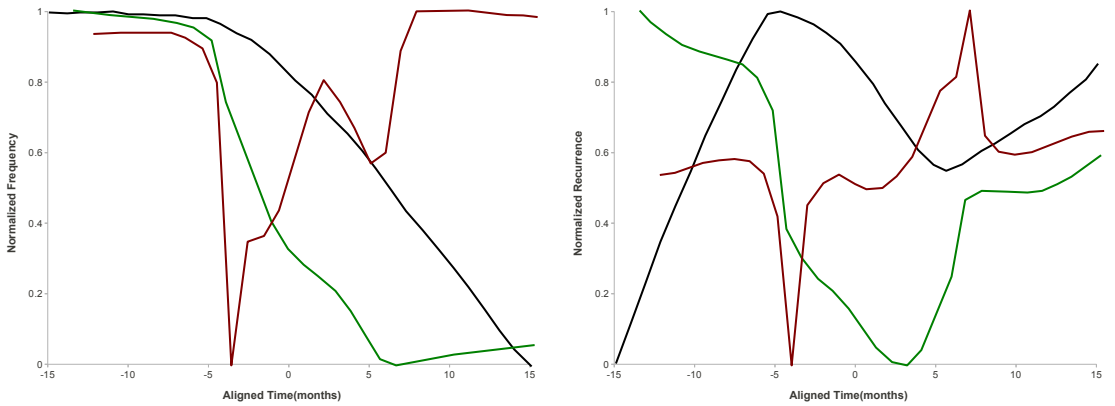
Table 6.2: Adaption scores and correlations with AoA for each of the predictor variables for each caregiver.

Caregiver 1		Freq.	Recur.	1/MLU	Dur.	F0	Int.
	Adaption Score	.57	.70	.86	.66	.60	.81
	Correlation with AoA	-.15	-.28	-.21	-.28	-.12	-.41
Caregiver 2							
	Adaption Score	.61	.60	.87	.72	.63	.77
	Correlation with AoA	-.18	-.29	-.24	-.19	-.20	-.31
Caregiver 3							
	Adaption Score	.70	.72	.92	.69	.50	.72
	Correlation with AoA	-.27	-.30	-.29	-.20	-.18	-.33
All							
	Adaption Score	.68	.73	.89	.77	.60	.80
	Correlation with AoA	-.23	-.37	-.25	-.29	-.19	-.35

6.3 Second Derivative Analysis

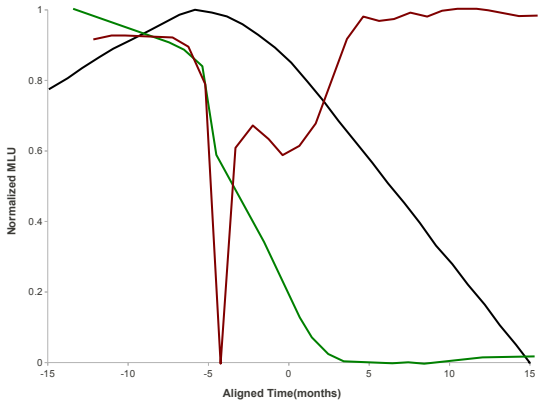
Looking at the original mutual influence curves for each variable 6-1, it seems that they all share a similar structure. In order to better visualize this shared structure, we looked at the first and second derivative of these curves as shown in Figure 6-5. The second derivative curves showed something surprising, a sudden valley appears in all the variables somewhere between 4 to 5 months before AoA.

We do not know exactly what is causing these sudden valleys though we have a theory that this sudden change in caregiver behavior, a few months before the AoA, might be happening around the age of comprehension (AoC) of the word. In other words, the caregivers change their linguistic behavior when mentioning a particular word when they believe they child has comprehended that word. However, the fact that this dramatic change in behavior happens around 120-150 days before AoA makes the case for AoC weaker since that might be too big of a gap between AoC and AoA. In the future work section we will go over possible studies that might help us understand exactly what is causing this sudden change in the linguistic behavior of the caregivers.

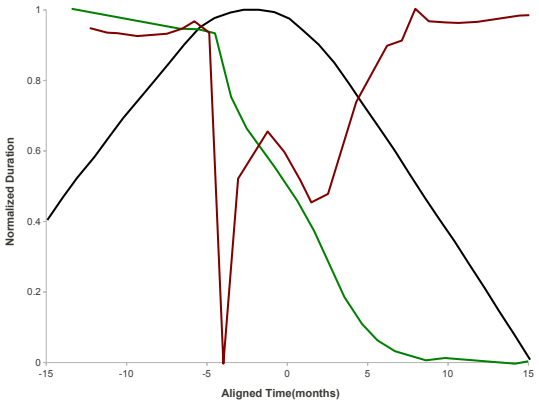


(a) Frequency

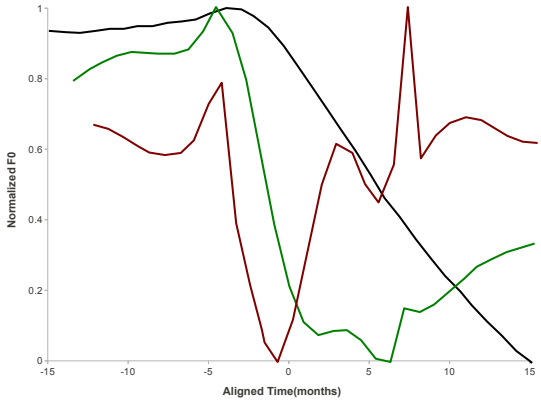
(b) Recurrence



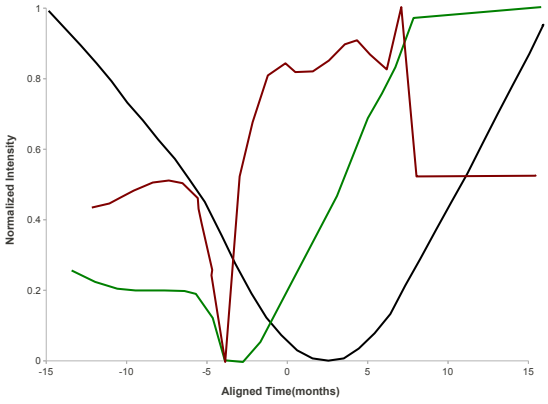
(c) MLU



(d) Duration



(e) F0



(f) Intensity

Figure 6-5: First and second derivatives of the mutual influence curves. The original curves are in black, the first derivatives are in green and the second derivative are in red.

6.3.1 Valley Detector

In order to better study this valley-effect at the word level we needed a way to automatically detect valleys similar to the ones we have observed in any given curve. To achieve this we create an automatic valley detector. The valley detector works by looking for the lowest point around which there are points higher by some value X (which we call the valley threshold) on both sides. We set the valley threshold by training our valley detector on positive and negative examples of what we considered a valley. The best performance for the valley detector was achieved when the valley threshold was set to be about 0.6. The valley detector also utilizes the fact that the valleys that we are interested in usually span across a month. The wideness of the valleys is used by the detector to reduce the number of false positives. The recall and precision rates for our valley detector were 0.97 and 0.93 respectively.

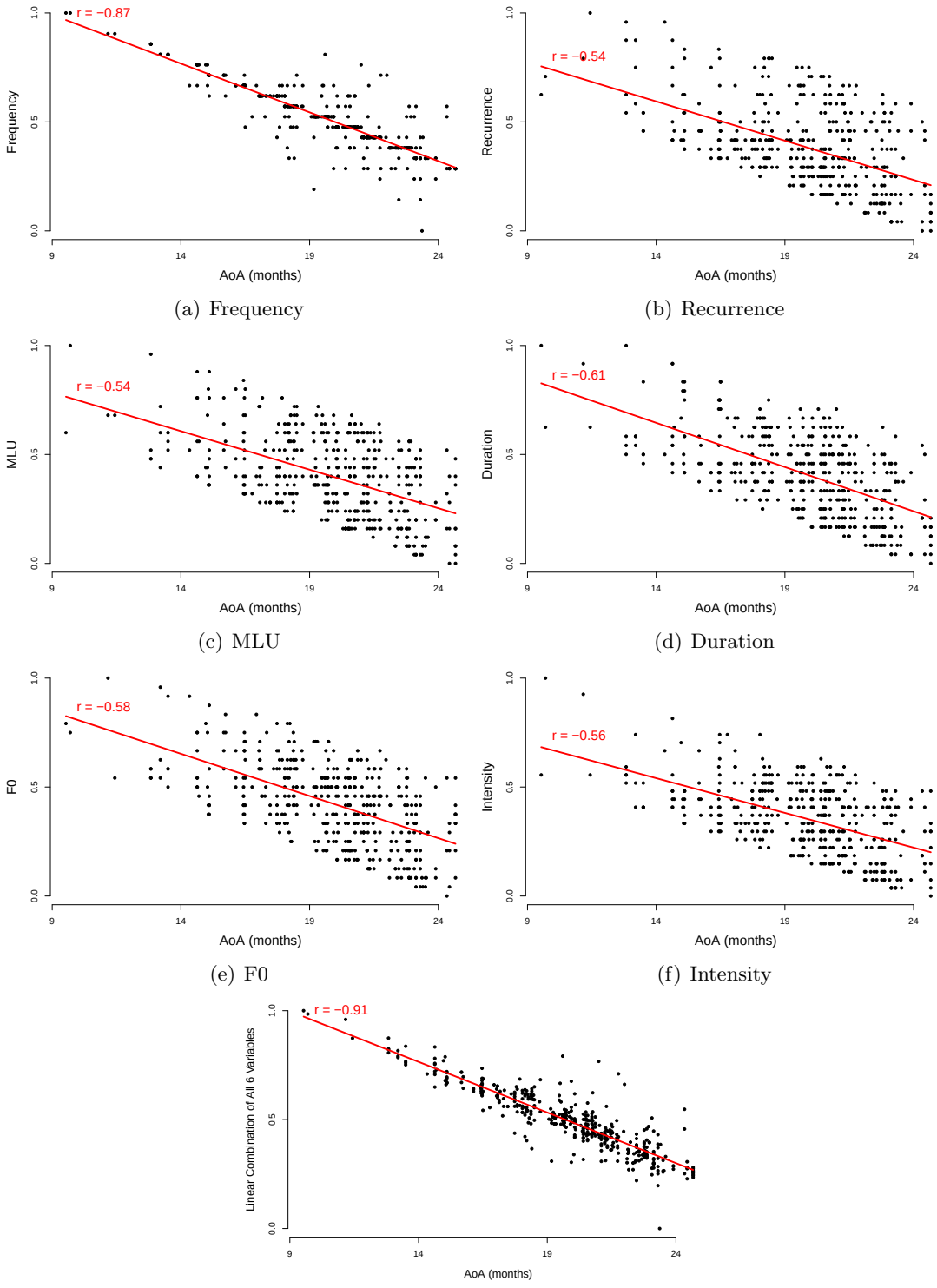
Chapter 7

Predictive Model 2

Given the unusually consistence appearance of a deep valley between 120-150 days before the AoA for all of the six variables that we coded in CAS, we next constructed a regression model which attempted to predict the AoA as a function of a linear combination of the day of the appearance of the valleys in the mutual influence curve of each of our 6 variables. Table 7.1 shows the correlations between AoA and the detected valleys in the second derivatives of the mutual influence curves of each of the 6 variables for each word category in the child's speech. All the correlations were negative, very strong and and highly significant (all p-values less than 0.001). The full linear model had $r = -0.91$, suggesting that it captured almost all of the variance in age of acquisition. Though the underlying cause of this strong correlation will require further study, it provides evidence of a new kind for fine-grained adaptive behavior by the caregivers in the context of child language development. The scatter plots of AoA vs the detected second derivative valleys for each of the six variables and their best linear combination can be see in Figure 7-1.

7.1 Evaluation of the Predictive Model 2

In order to evaluate the fitness of our new predictive model we used the same k-fold cross validation (K=461) technique that was used to evaluate our first model. To recap, this was



(a) Frequency

(b) Recurrence

(c) MLU

(d) Duration

(e) F0

(f) Intensity

(g) Best Linear Combination of All 6 Variables

Figure 7-1: Each subplot shows the univariate correlation between AoA the detected valley in the second derivative of the mutual influence curve of a particular predictor variable. Each point is a single word, while lines show best linear fit.

Table 7.1: Pearson's r values measuring the correlations between age of acquisition and the age of the second derivative valleys for each category in child's speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adjectives	Nouns	Verbs	All
Frequency	-.89**	-.79**	-.98**	-.87**
Recurrence	-.48**	-.45**	-.39**	-.54**
1/MLU	-.48**	-.41**	-.46**	-.54**
Duration	-.49**	-.58**	-.58**	-.61**
F0	-.50**	-.62**	-.48**	-.58**
Intensity	-.46**	-.46**	-.38**	-.56**
All combined	-.90**	-.85**	-.98**	-.91**

done by ignoring one of the 461 words in the child's lexicon and redoing our entire analysis- from variable optimization to the second derivative analysis and regression analysis- on the remaining 460 words. This way the new model is completely blind to the the word that was left out. We then try to predict the age of acquisition of that particular word with our new model. Figure 7-2 shows how this is done. First, we generate the mutual influence curve for all six predictor variables using all CAS utterances that contain the word whose age of acquisition we are trying to predict. Note that we do this for utterances spanning the whole 9-24 months timespan (as opposed to utterances up to AoA) since the predictor is supposed to be blind to the AoA of the word whose AoA the model is trying to predict. Next, the second derivatives of the mutual influence curves are generated which are then run through the valley-detector which returns the day at which the valleys appear. These days are then used as inputs for our predictive model which then in return predicts the age at which the word in question will be acquired. We do this for all the 461 words in our corpus. We can then compare the true age of acquisition of the 461 words with the predicted age of acquisition.

Figure 7-3 shows the relation between predicted age of acquisition (via the new predictive model) and the true age of acquisition of words by the child. A perfect model (with a correlation of 1.0) would generate a plot with a straight line at a 45 degree angle (since all the predicted and true AoAs would match). Our new model (with a correlation of 0.91) is getting very close to a perfect model as evident by Figure 7-3. On average our new model

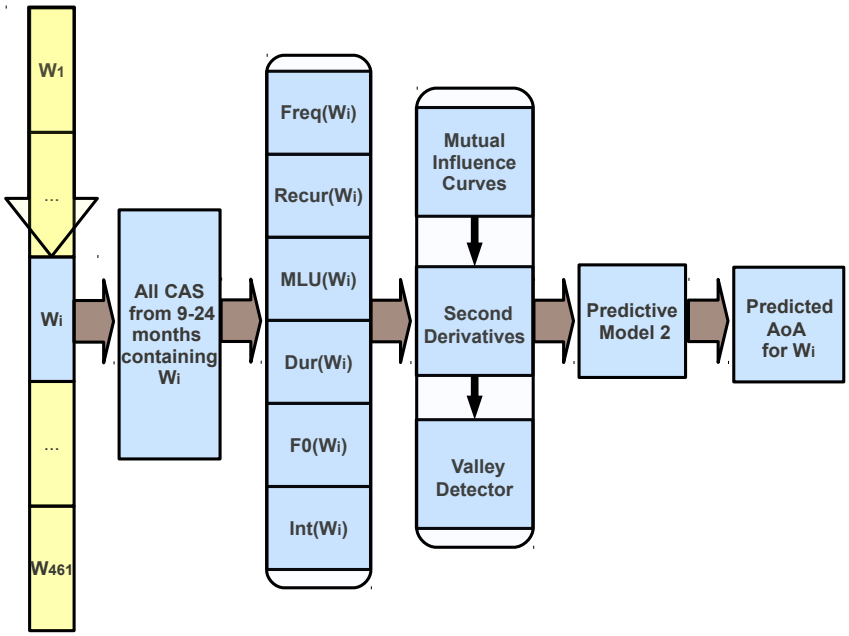


Figure 7-2: Schematic of the pipeline used for the 461-fold cross validation of predictive model 2.

can correctly predict the age of acquisition of a word by the child within 10 days, compare that with our previous model which was able to correctly predict the age of acquisition of a word by the child within 55 days

These very strong results make us hesitant to claim with utter certainty that there is indeed prediction happening, though that might very well be the case. As discussed in the previous section, the valleys in the second derivatives (which are ultimately used in this model for prediction) might be indicating the change in caregiver linguist behavior at around the age of comprehension of a word (we will go over possible studies to test this hypothesis in the future work section), in which case prediction might not be the correct term to use for what we have done here. Whatever the case may be, these results are so strong and shocking that they deserve further analysis and study.

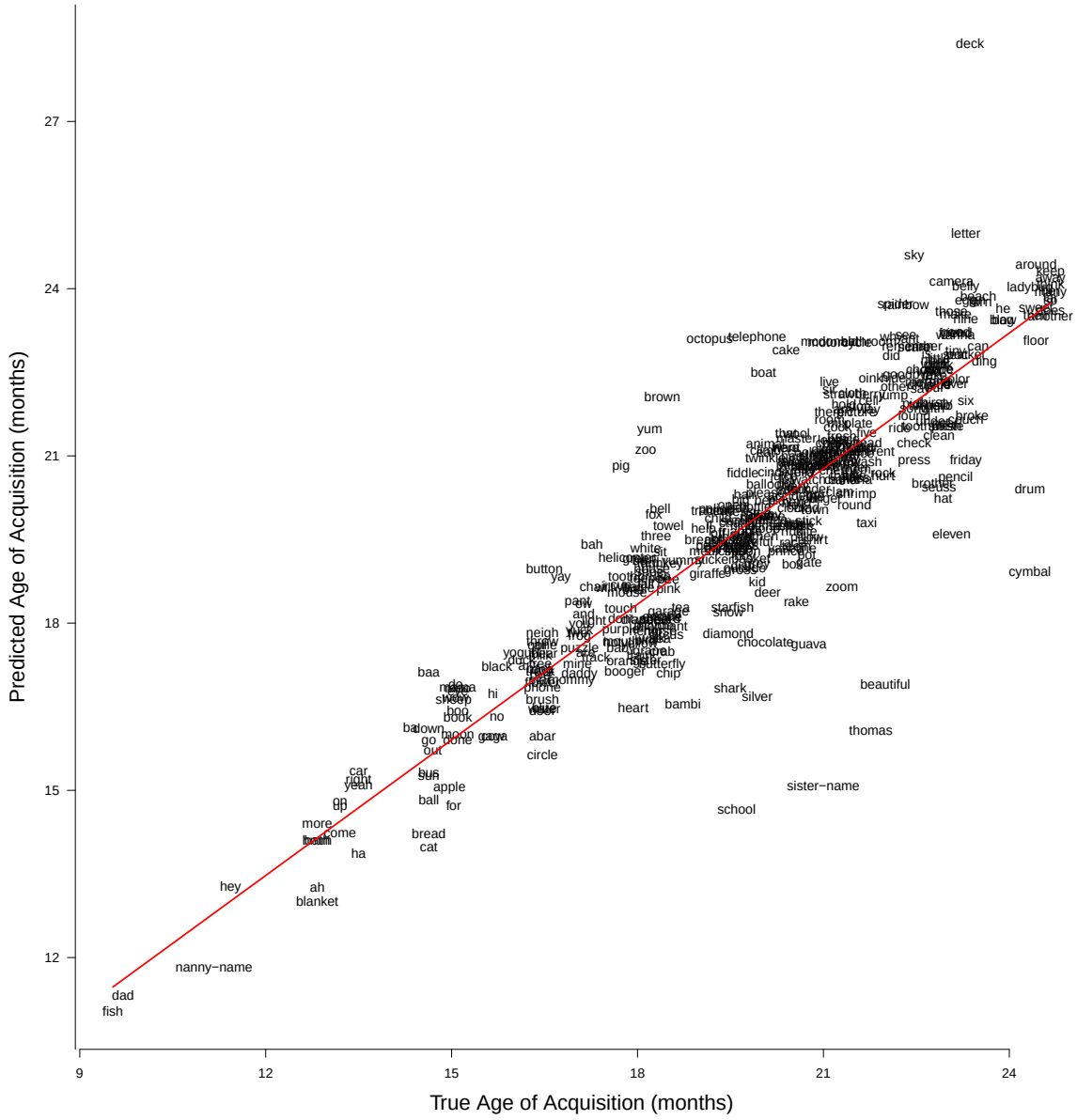


Figure 7-3: Predicted AoA by model 2 vs. true AoA

Chapter 8

Child Directed Speech: A Preliminary Analysis

As mentioned in the methods section of this thesis, the corpus that was used for the analysis described so far consists of child available speech (which we defined as caregiver speech when the child is awake and close enough to hear) as opposed to child directed speech(CDS). Part of the reason that we did not consider CDS was that the annotation of CDS requires a great amount of human effort.

In order to remedy this problem, we developed an automatic child directed speech detector. We then used this detector to identify CDS in our corpus. Next, we conducted a preliminary study of the effects of CDS on the performance of the models developed so far in this thesis.

8.1 Child Directed Speech Detector

The main reason for postponing the analysis of CDS was the sheer effort that it would take for humans to manually identify CDS. Therefore, the first thing that needed to be done before we could start using and analyzing CDS, was to come up with a classification algorithm to automatically differentiate CDS from non-CDS.

8.1.1 Corpus Collection

In order to train and test our classifier we needed to collect positive and negative examples of child directed speech as our ground truth. To achieve this we created a simple tool with a very simple interface to be used by humans to collect samples of CDS(Figure 8-1). The tool plays random (uniformly distributed over time) audio clips of child available speech along with their transcripts. The transcribers can then choose between "CDS" and "other" to manually classify the audio segment.

A total of 4250 audio segments distributed over 9-24 months were manually annotated by humans.

Figure 8-1: Tool used by human annotators to generated ground truth for the CDS detector.

8.1.2 Features

We considered many different acoustic, linguist and other features to be used in our CDS detector. Out of all the features a few seemed to be extremely good at distinguishing child directed and non-child directed speech. These features are: duration of phonemes, intensity, fundamental frequency (F0), time-of-day and length of utterance (LU). These features agree with previous studies that have shown that CDS has special characteristics such as shorter

utterance lengths and exaggerated prosody [21]. As an added benefit, all of these features have already been extracted and optimized for the entire corpus (described fully in the methods section of this thesis).

These features make intuitive sense as well. Duration, intensity and fundamental frequency are a proxy for prosody, which would be expected to be different in child directed speech vs adult directed speech. Moreover, LU is an estimate of the linguistic complexity of an utterance, and it would be expected for child directed speech to be linguistically simpler than adult directed speech. Finally, there are certain times of day when the caregivers are more likely to be interacting with the child. For example it is very unlikely that speech at 2am is directed at the child; the time of day feature captures this aspect of the data.

In addition to these five features, we also constructed two bi-gram language models using HTK [38], one for child directed speech and one for non-child directed speech. We created these two language models for all three primary caregivers, giving us a total of six different language models (3 speakers, 2 categories). Given a sample utterance from a caregiver, using these models we can calculate the probability of that utterance belonging to the child-directed language model and the non-child directed language model. Therefore, all together there is a total of seven features used by our CDS detector as shown in Figure 8-2.

8.1.3 System Architecture

Figure 8-2 shows the pipeline used in the classifier. As mentioned, duration, F0, intensity, LU and time-of-day all have been already extracted and optimized for our dataset(since they were needed for our analysis so far) so they are readily available. The classifier used here is a standard boosted decision tree.

8.1.4 Evaluation

The child directed speech detector was evaluated using 10-fold cross validation using our corpus. The recall and precision rates for the CDS detector were 0.91 and 0.86 respectively.

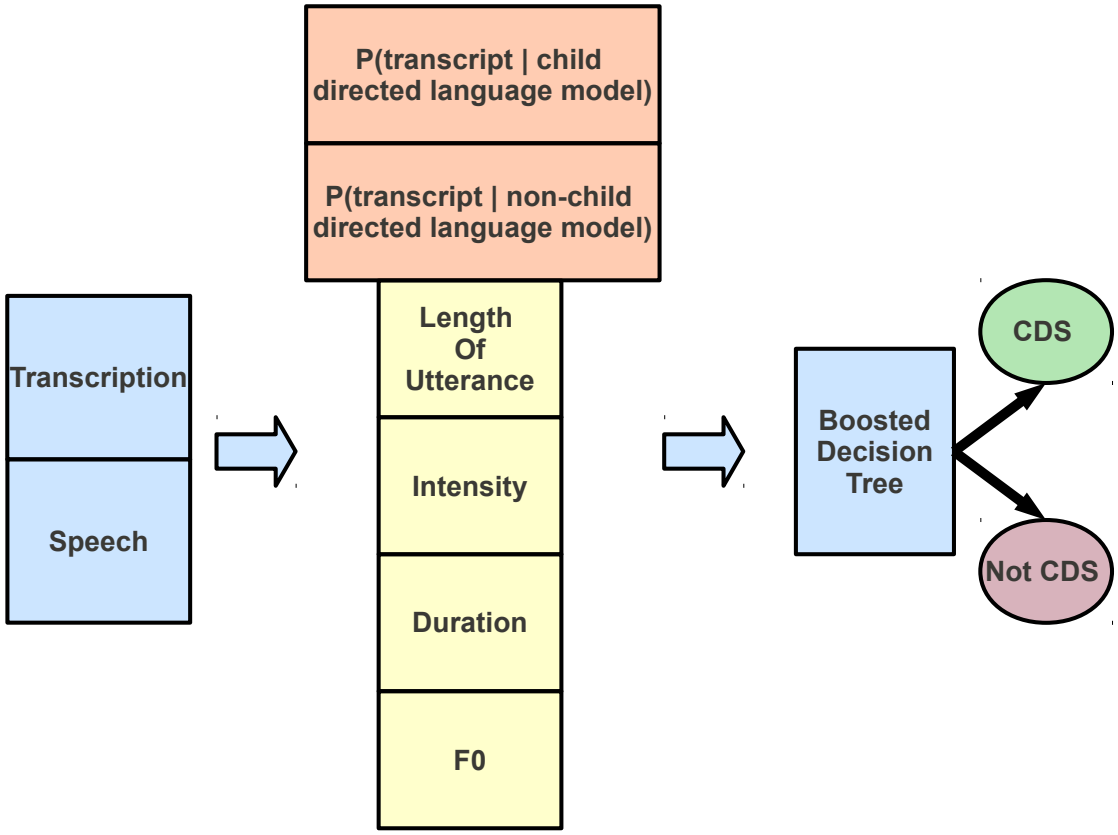


Figure 8-2: Schematic of the pipeline for classification of child directed speech. The yellow boxes are features that have already been defined and extracted for use in other parts of this thesis. The orange boxes are new features that have been defined specifically for the CDS classifier. The boosted decision tree is a binary classifier which classifies the speech as either CDS or not-CDS.

Understandably, the system was less reliable when classifying speech from anyone other than the three primary caregivers for whom we developed child directed speech language models.

8.2 Child Directed Speech vs Child Available Speech

With the CDS detector trained and evaluated, we next ran our entire corpus through the detector (again utilizing our parallelized infrastructure) and labeled all child directed speech. Next we redid our entire analysis (described so far in this thesis) on CDS in place of CAS which we did our original analysis on. Table 8.1 shows the difference in performance between CAS and CDS for both of our models. The use of CDS significantly increased the model fit for both of our models, though it had the greatest effect on our first model. As shown in 8.1, the correlation for the fist model improved from -.55 to -.64 while the correlation for the second model improved from -.91 to -.95 (all improvements where statistically significant).

Table 8.1: Pearson's r values measuring the fitness of our two predictive models running on CAS vs CDS. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Model 1	Model 2
CAS	-.55**	-.91**
CDS	-.64**	-.95**

Our models become more predictive of age of acquisition of words by the child when using CDS vs CAS. This result has very interesting implications as to the utilization of CDS vs CAS by the child in word acquisition. Due to time constraints we were only able to look at the overall performance of each model using CDS, we did not have time to look closely at the difference in performance of CDS vs CAS for different predictor variables and word categories. Therefore, these results should be at the very best be treated as a preliminary study of affects of CDS on word acquisition. In the future work section of this thesis we will discuss possible next steps to continue this line of analysis.

Chapter 9

Fully Automatic Analysis

Part of the motivation of this thesis was to develop models for child word acquisition in order to utilize them in artificial language learning systems. The models created so far serve this purpose well with one caveat, the models are not fully automatic. For these models to be implemented and utilized by artificial systems, they need to be fully automatic. In this chapter, we develop and evaluate a fully automatic predictive model. This model is capable of predicting the AoA of words by the child by processing and analyzing raw audio from the HSP audio corpus without any human intervention.

Figure 9-1 shows the complete processing pipeline used in this thesis, from raw audio to correlation analysis and predictive model. The green boxes represent automatic components while the red boxes represent non-automatic or semi-automatic components. Except for the audio transcripts, everything in our processing and analysis pipeline from speech detection to variable extraction and optimization to our k-fold cross validation is done by automatic systems. Therefore, in order to have a fully automatic system capable of predicting AoA of words from raw audio data we would need to have an automatic transcription system. In other words, we need an automatic speech recognizer for the Human Speechome Project.

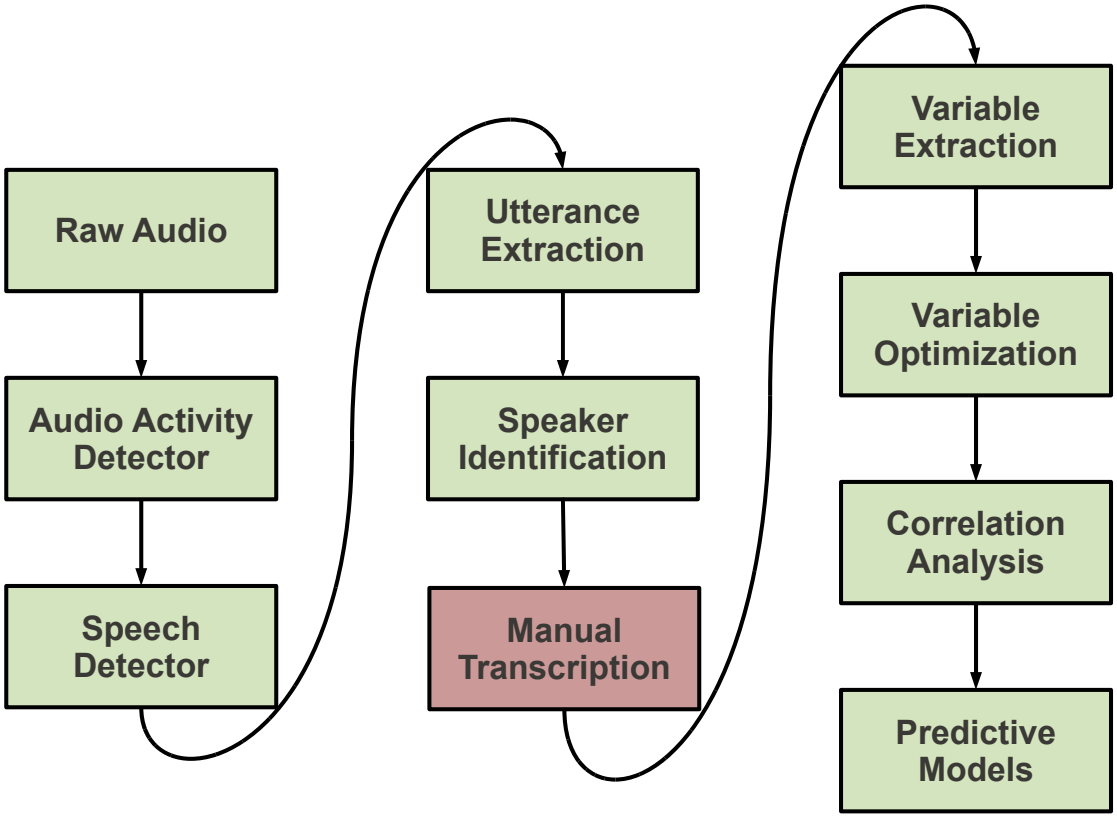


Figure 9-1: Overview of the processing pipeline used to get from raw audio to the analysis that was done in this thesis. The green boxes represent automatic components while the red boxes represent non-automatic or semi-automatic components.

9.1 Automatic Speech Recognizer for HSP

Given the sheer volume of our transcribed audio corpus, it was relatively easy to train an automatic speech recognizer (ASR) for our corpus using HTK [38]. In order to train our ASR using the HTK, we needed to first train acoustic and language models for our dataset. We trained four distinct acoustic and language models, one for each of the three caregivers and one for the child. The next two subsections describe how these models were trained.

9.1.1 Speaker Dependent Acoustic Models

The caregiver acoustic models used here are the same ones developed for the forced-aligner in section 2.3.5.1.1. An additional acoustic model was developed for the child using HTK. The acoustic model created for the child is not very strong as the acoustic characteristics of the child's voice changes dramatically as he gets older. We will discuss possible solutions to this in the future work section of this thesis.

9.1.2 Speaker Dependent Language Models

HTK was also used to create four distinct bi-gram language models from the transcripts, one for each speaker and one for the child. As with the acoustic model, the language model created for the child is not very strong as the child's linguistic complexity evolves over time.

9.1.3 Evaluation

The evaluation process for our ASR was fairly simple since we already had thousands of transcribed audio segments available in our corpus to be used for testing our ASR.

We ran a total of 1600 audio segments through our ASR (500 for each speaker and 100 for the child). We rated the accuracy of the recorded transcription for each segment with word error rate (WER). As shown in Equation (9.1) WER is computed by aligning the recognized

word sequence with the reference(from our human annotated transcription) word sequence. In Equation (9.1), S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the reference.

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}} \tag{9.1}$$

For our evaluation we look at word accuracy (WAcc) which is 1 minus WER, as shown in Equation (9.2).

$$\text{WAcc} = 1 - \text{WER} = \frac{\text{N} - \text{S} - \text{D} - \text{I}}{\text{N}} \tag{9.2}$$

As might be expected, the ASR performs very poorly on some transcripts and strongly on others. Figure 9-2 shows the accuracy of the ASR for different speakers vs yield. Yield is the measure of how much of the data we keep and how much we throw out. For example, a 20 percent yield means that only the top fifth (in terms of accuracy score) of the automatically transcribed utterances are kept while the bottom 80 percent are been thrown out. As expected, ASR's performs very poorly almost on all child speech. We set the cutoff for our ASR at 20% yield.

9.2 Automatic Analysis: First Pass

With the ASR trained and evaluated, we next ran our entire audio speech corpus through the ASR (again utilizing our parallelized infrastructure) and automatically generated transcripts for the speech. Next we redid our entire analysis on the automatically generated transcripts. Table 9.1 shows the difference in performance between the models constructed using human generated transcripts versus automatically generated transcripts. The automatic models' performances are extremely weak($r = -.07, p = 0.49$ and $r = -.10, p = 0.42$). In fact non of the models are statistically significant.

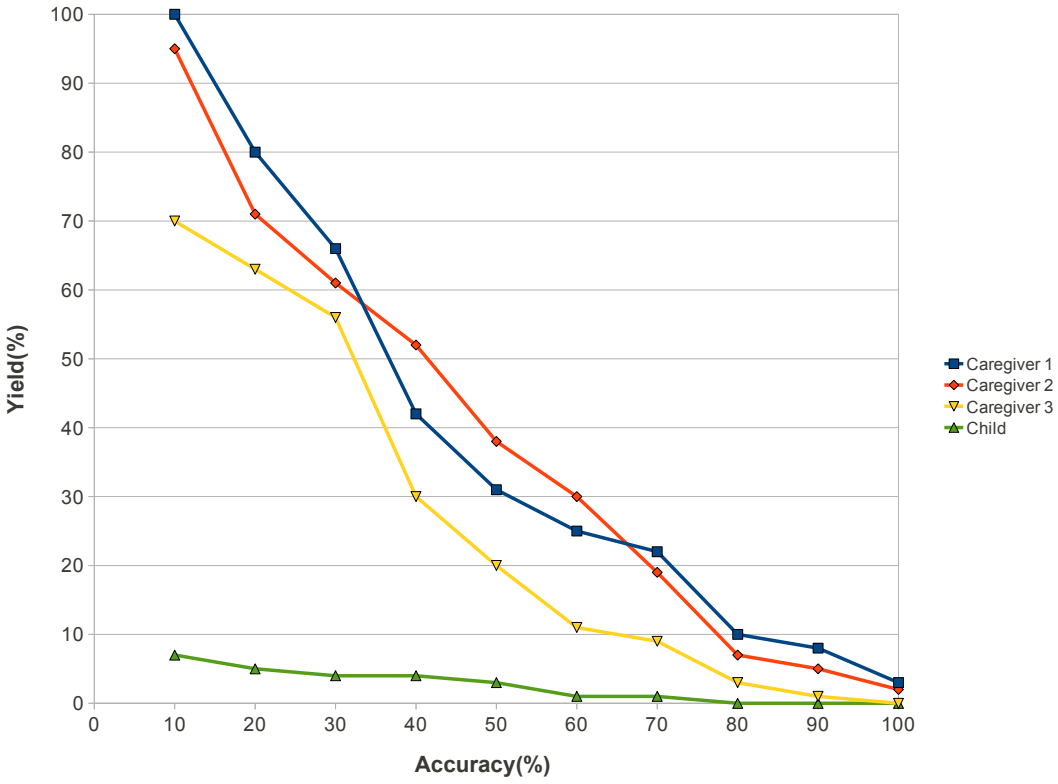


Figure 9-2: Accuracy of the ASR for different speakers in the HSP corpus.

Table 9.1: Pearson's r values measuring the fitness of our two predictive models running on human transcribed vs automatically transcribed audio. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Model 1	Model 2
Human Transcribed	-.55**	-.91**
Automatically Transcribed	-.08	-.11

One of the biggest chokes in our automatic analysis is the very poor performance of the ASR on child speech. One of the core components of our analysis is the age of acquisition of words by the child, which we defined as the first recorded used of the word by the child. If the ASR can not correctly identify the words uttered by the child then we cannot correctly identify AoA of words. Without accurate AoA for the words, the whole analysis falls to pieces as shown in Table 9.1.

9.3 Automatic Analysis: Second Pass

We reran our automatic analysis a second time but this time only focused on using the automatically generated transcripts for the caregivers and not the child. We used the manually annotated AoAs of the words in the child's lexicon. Like before, we redid our entire analysis. Table 9.2 shows the difference in performance between the models constructed through all three approaches (human generated transcripts, fully-automatic transcripts, automatic caregiver transcripts). Though still statistically not as significant as our original analysis, the models generated by this method were significantly more fit than the models generated by the previous methods($r = -.13, p = 0.14$ and $r = -.28, p = 0.09$).

Table 9.2: Pearson's r values measuring the fitness of our two predictive models running on human transcribed vs automatically transcribed audio excluding child speech. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Model 1	Model 2
Human Transcribed	-.55**	-.91**
Automatically Transcribed	-.08	-.11
Automatically Transcribed(excluding child speech)	-.13	-.28'

Chapter 10

On-line Prediction of AoA

The predictive models that have been developed so far in this thesis were able to predict the AoA of words learned by the child from 9–24 months by processing and analyzing six predictor variables coded in CAS (or CDS in section 8) in that 9-24 months timespan. In other words, our predictive models had access to the whole 9–24 months corpus and were using that to do an "off-line" prediction of when the child learned the words that he did.

However, in order for our models to be a more valid representation of the mechanisms used by the child for word learning, and to maybe at a later point be implemented and used by artificial language learning systems grounded in the real world, they need to be able to do "on-line" prediction of AoA. In other words, our models should be able to take the place of the child in the audio corpus, listen to all the speech that the child had access to in a chronological fashion and while listening to the audio calculate the probability of all words(individually) having been learned by the child by that point in time. In this chapter we go over the design and performance of one such on-line prediction system (shown in Figure 10-1).

10.1 System Architecture and Design

To implement an on-line predictive model, we used our second model described in Chapter 7 of this document. Recall this model predicts the AoA of a word by detecting the valleys in the second derivatives of the mutual influence curves (described in Chapter 6) of that word and using the days at which the valleys were detected as predictors for the AoA of the word in question. The way this model operates makes it perfect for doing on-line prediction of AoA.

Figure 10-1 shows the processing pipeline of the on-line prediction system. The system is constantly being fed CAS speech and transcriptions from the corpus in chronological order. At the end of each day (for simplicity we discretized time into units of day) the system generates mutual influence curves for all the six variables for all words using the accumulated CAS from the beginning of the corpus to the end of that day. The system then generates the second derivative of these curves. Next, the system runs these curves through the valley detector (described in section 6.3.1) in order to detect the valleys that appear a few months before AoA. The very low number of false positives generated by our valley detector (mostly due to the unusual wideness of the valleys that we are interested in) and also the fact that our system is looking for valleys in the mutual influence curves in all six variables (not just one) ensures that the on-line prediction system is not lead astray by an early false-valley detection.

After all the computation is done, the system then can estimate the probability of the child having already acquired a particular word by that day. Usually 2 months before the AoA of a word the valleys are fully visible and detectable for all the 6 variables. Since AoA usually happens at least 2 months ahead of these valleys, the system keeps the probability of a word having been acquired by the child very low (almost close to zero) until about 2 months after the appearance of the first valley. The system rapidly increases the probability of a word having been acquired by the child when supported by valleys from other variables (the probability increases as the agreement between the variables increases).

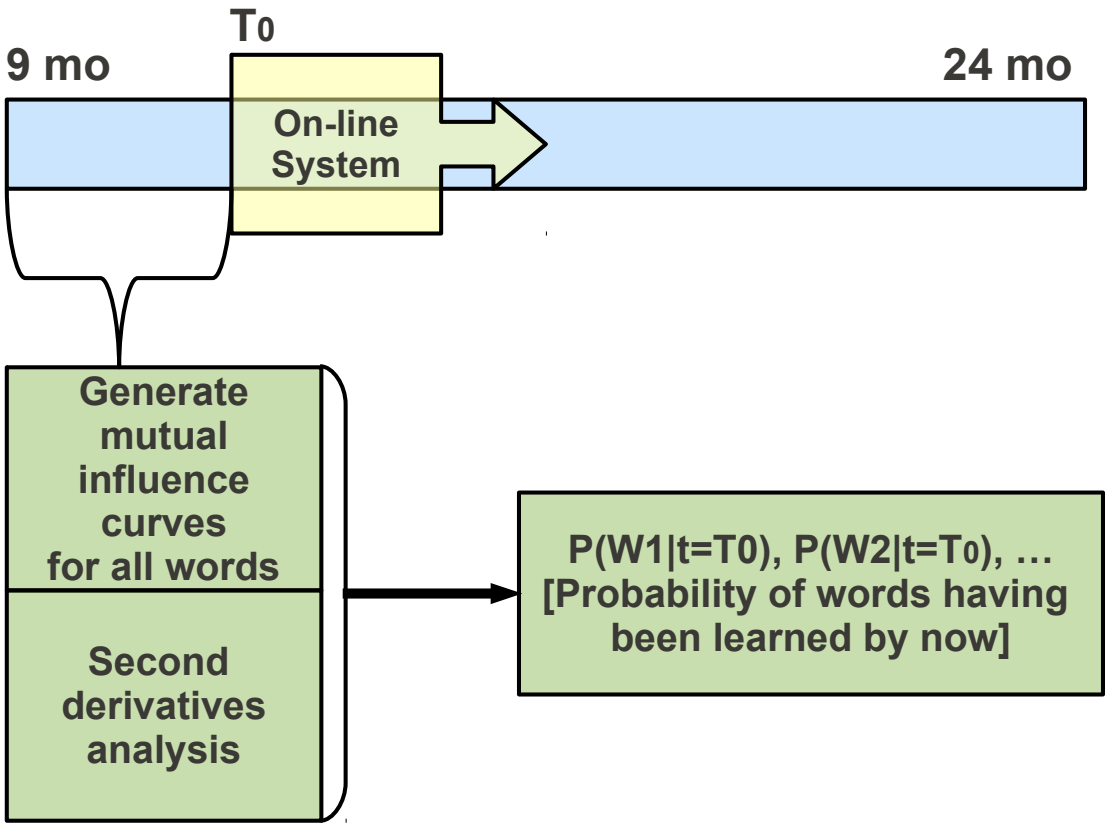


Figure 10-1: Processing pipeline of the on-line prediction system.

10.2 Results

Figure 10-2 shows the average result of on-line prediction done on all the 461 words. The x-axis on the graph is time, ranged from 15 months before AoA to 15 months after AoA. This is done because in the 9–24 months range, words with AoAs on month 9 have 15 months ahead of the AoA while words with AoAs on month 24 have 15 months before the AoA. The y-axis is the probability that a word has been acquired by the child at a given day. The probability is almost 0 until about 2 months before the actual AoA. The probability then slowly increases until about 15 days before the AoA after which it rapidly increases until about 10 days after the AoA at which point its growth again slows down as it asymptotically reaches 1.

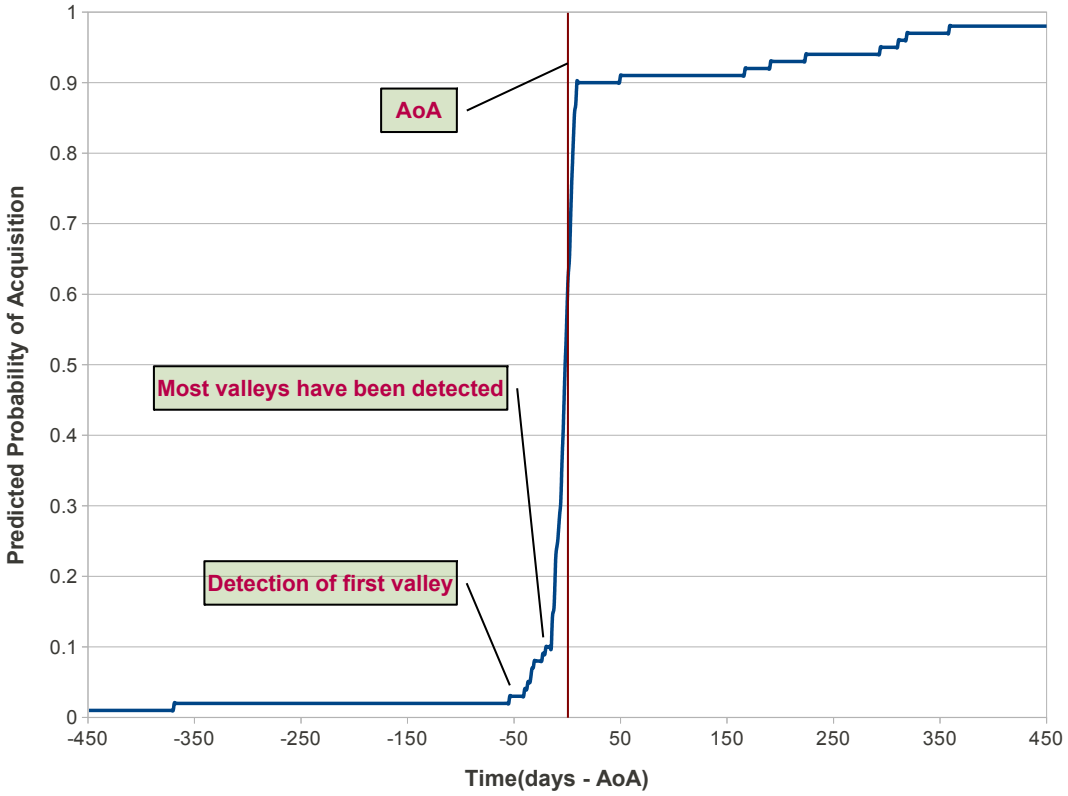


Figure 10-2: The averaged output of the on-line prediction system running on all 461word.

Chapter 11

Contributions

Using independent and creative methods of data analysis and through a series of computational studies and explorations on the fined-grained interactions of caregiver speech and one child's early linguistic development, this thesis helps us better understand the relationship between caregiver speech and early word learning and helps illuminate the underlying mechanism in child language acquisition.

As part of the thesis, we create a collection of tools integrated into a software analysis environment for fast processing and analyzes of high density, longitudinal corpora such as the Human Speechome Corpus.

Moreover. this thesis explored the relationship between a single child's vocabulary growth and prosodic and distributional features of the naturally occurring caregiver speech that the child was exposed to. We measured F0, intensity, phoneme duration, usage frequency, recurrence and MLU for caregivers' production of each word that the child learned from 9-24 months. We found significant correlations between all 6 variables and age of acquisition(AoA) for individual words, with the best linear combination of these variables producing a correlation of $r = -.55 (p < .001)$. We then used these variables to obtain a model of word acquisition as a function of caregiver input speech. This model was able to accurately predict the AoA of individual words within 55 days of their true AoA.

Furthermore, this thesis looked at the fine-grained temporal relationships between caregivers' speech and the child's lexical development. We saw significant evidence of caregiver tuning for all of the 6 variables that we coded in caregiver speech. The tuning behavior was remarkably consistent across caregivers and variables, all following a very similar pattern. We found significant correlations between the patterns of change in caregiver behavior for each of the 6 variables and the AoA for individual words, with their best linear combination producing a correlation of $r = -.91 (p < .001)$. Though we are not sure what this strong correlation captures, it is strong evidence for fined-grained adaptive behavior by the caregivers in context of child language development. We then used these patterns to obtain a model of word acquisition as a function of change in caregiver behavior for each of the 6 variables. This model was able to accurately predict the AoA of individual words within 10 days of their true AoA.

Additionally, this thesi provided a preliminary analysis of child directed speech vs child available speech(made possible by the development of an automatic child directed speech detector). We showed statistically significant improvements to our models ($r = -.64$ and $r = -.96$ for the first and second models respectively) when using child directed speech in place of child available speech. In other words, our models become more predictive of age of acquisition of words by the child when using CDS vs CAS.

Moreover, in this thesis we developed and evaluated a fully automatic system which attempted to automatically replicate all studies done in the thesis from raw audio in the corpus without any human processing or transcription. The system was able to get statistically significant results ($r = -.28, p < .1$) when replicating the second model which uses change in caregiver behavior as a predictor for AoA. Such automatic systems will be invaluable when developing artificial language learning systems.

Finally, this thesis saw the development of an on-line prediction system capable of calculating the probability of words beings learned by the child on-line, while listening to the audio corpus. Such on-line prediction systems also open the door to using our models in artificial language learning systems grounded in the real world.

Chapter 12

Future Work

In addition to the work covered in this thesis, the are many other interesting research questions that can be investigated using the extremely rich HSP dataset. Moreover, we have plans to utilize a new recording infrastructure which we call the *Speechome Recorder (SHR)* in order to collect data similar to that of HSP from several families across the US. In this section we will go over possible next steps in the Human Speechome Project and the possibilities of our new recording technology, the Speechome Recorder.

12.1 Human Speechome Project

There are a few interesting research questions involving the corpus of the Human Speechome Project that have been brought up throughout this thesis that need further study and analysis. Some of these future research threads can be studied and accomplished in a few short months while others might take longer. Inspired by that fact, this section of the thesis is divided up into short term and long term research goals.

12.1.1 Short Term

The short term goals are mostly improvements on our current analysis and computation methods. Here we will go over three such goals.

12.1.1.1 More Detailed Study of CDS

In Chapter 8 of this thesis we briefly looked at the performance of our models using CDS. We showed our predictive models getting stronger when running on CDS vs CAS. However, we did not have time to fully study the difference in the performance of our models on CDS vs CAS. We need to look at the performance of CDS vs CAS for different POS and CDI categories. We might find the importance of CDS in word acquisition to vary across different word categories. Though unlikely, it would be interesting if we identified certain word categories where CAS has a more important role to play than CDS in child word acquisition.

Furthermore, we need to look at the performance of CDS vs CAS for each caregiver. It is very likely that for some caregivers, for example the nanny, the difference in performance between CDS vs CAS would be insignificant since almost all of the nanny speech captured in our corpus is directed at the child.

In brief, there is a great opportunity here to study the importance of CDS in child word acquisition.

12.1.1.2 Managing Inaccurate Transcriptions

Even with human transcribers, there is chance of inaccurate transcriptions. Though a majority of our transcripts are accurate, there is still a noticeable percentage of inaccurate transcripts. These inaccurate transcripts get used in our analysis along with the accurate transcripts, no doubt compromising our analysis and hurting the performance of our models.

Measuring transcription accuracy by calculating inter-transcriber accuracy scores is very costly in a dataset as large as the HSP, as each audio segment has to be transcribed by multiple transcribers. That is why we developed a system that can automatically estimate the accuracy of a transcription [29]. The system has been evaluated and is ready to be used. In the not-distant future we will utilize this tool to automatically drop inaccurate transcriptions before doing our analysis.

12.1.1.3 Time-dependent Child Acoustic and Language Models

As discussed briefly in section 9.1.1 and 9.1.2 of this thesis, the acoustic and language models trained for the child are extremely ineffective and almost unusable. The main reason behind this is that the child's acoustic and linguistic characteristics changes as the child gets older, almost from month to month. The solution to this problem is simply to use time-dependent acoustic and language models for the child. We can easily do this by training separate acoustic and language models for the child for each month from 9-24 months of age. Our system can then decide which acoustic and language models to use based on the month in which it is operating.

12.1.2 Long Term

The long term goals are research questions that have come to surface in the span of our thesis. Each of these research threads is very rich and deserves a much closer study.

12.1.2.1 Study of Outliers in Our Models

In section 4.2 we briefly discussed the outliers in our predictive model. Words that deviate from our model's predictions are of interest as they may suggest important social, contextual and other cures relating to word learning that our models are missing. We would like to further probe the outliers in our models to better understand these cues.

For example in Figure 4-3, the words fish, cat and bird (which are matched int terms of semantic complexity) are predicted by model 1 to be acquired by the child at about the same age (17 months); however, in actuality the child learns these words at 10,15 and 19 months of age respectively. Moreover, words like dad and fish are predicted by our model to be acquired by the child months away from each other (at 17 and 21 months respectively), while the child actually learns those words at around the same age (10 months).

In order to study these outliers in greater detail, we need to manually investigate video clips of episodes where these words were used by caregivers and annotate visual and social cues and compares these cues with those of similar words that are not outliers. This study at the very least can help us better understand the importance and role of social and visual cues in child word acquisition. At the very best, this study has the potential to illuminate some previously unknown factors that affect child word acquisition. The manual component of this study makes it a long term study since it would take many hours to manually watch and annotate the episodes containing even one word.

12.1.2.2 Further Study of Second Derivatives of the Mutual Influence Curves

The unusually strong predictive power of the valleys in the second derivatives of the mutual influence curves of all of our variables (discussed in section 7.1) warrants further study. We would like know why the linguistic behavior of the caregivers with regards to a particular word changes so greatly 4 to 5 months before the AoA of that word.

One theory is that the valleys in the second derivative mark the age of comprehension of words by the child. The caregivers realizing that the child has started comprehending a word, then adapt and rapidly change their linguistic behavior around that word (e.g. using the word in simpler sentences) in order to further encourage the learning of that word. This theory however assumes that AoC and AoA of a word are almost always between 4-5 months apart, which seems not very likely.

In any case, in order to really understand the cause behind the change in caregiver behavior, we need to manually examine episodes of CAS at around the time when the valleys appear

in our graphs. Given a word, lets say "water", we need to calculate the mutual influence curves for "water", mark the valleys in the second derivatives and then manually examine episodes where the word water is being used in CAS around the time of the valleys. This should help us understand the core cause behind the change in caregivers' behavior.

As mentioned earlier in the thesis, AoA is a conservative estimate of AoC. Therefore, if through this study, it was confirmed that the valleys do indeed mark AoC, then we can redo the entire analysis done in this thesis replacing the AoA in our models with AoC to hopefully get a more accurate representation of word acquisition by the child.

12.1.2.3 Multi-modal Models

The models described in this thesis only take into account variables in caregiver speech, omitting the visual and social context of word learning. One of the benefits of the Speechome corpus is that this information is available through rich video recordings. Computer vision algorithms and new video annotation interfaces are being developed to incorporate this aspect of the corpus into future investigations. In addition, our work in this thesis has been limited to the child's lexical development; our plan is that future work will extend the current analysis to grammatical development.

12.2 Speechome Recorder

One of the major limitations of the work presented in this thesis is the fact that Human Speechome corpus, though extremely dense and longitudinal, only captures the linguistic development of one child. In order to expand the corpus to multiple children, a new recording tool called the Speechome Recorder has been developed(seen in Figure 12-1). The Speechome Recorder (SHR) is a self-contained and compact audio and video recording tool that can be easily installed in any house within minutes. We will describe the SHR in more detail in the design section below.

12.2.1 Design

The Speechome Recorder as shown in Figure 12-1 consists of a head with a fish-eye camera and a microphone and a body with a frontal camera and a touchscreen for the user. SHR is completely self-sufficient in terms of data capture and storage. The recording and storage software and format used for the SHR is identical to the one used for the HSP. Therefore, all the algorithms and processing pipelines developed for the HSP can also be used on data from the SHR. The SHR has the added benefit of having a frontal camera (which the HSP lacked) which enables us to capturing and analyze facial expressions.

12.2.2 Speechome Corpus: A Look Ahead

Looking ahead, the Speechome Recorder will allow us to collect longitudinal, dense and naturalistic datasets for multiple children. Though due to logistical limitations, the datasets generated by the SHR will most likely not be as dense as that of the Human Speechome Project.

On a final note, the longitudinal and naturalistic nature of the SHR, coupled with the fact that the SHR corpus will capture the linguistic development of multiple children across several households, make the SHR invaluable for studying autism and other child developmental disorders. The nature of datasets that will be generated by the SHR will allow us to study developmental trajectories of autistic children and compare them to those of typically developing children.

12.3 Final Words

All the work presented in this thesis focused on modeling the input-output relationships in child languages acquisition. I believe in order to truly understand the mechanisms underlying child language acquisition we need to understand and model the *p*rocesses with which the child acquires languages.

Figure 12-1: Prototype of the Speechome Recorder(SHR).

Bibliography

[1] Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 5.1.01). http://www.praat.org/, 2009.

[2] R.P. Cooper and R.N. Aslin. Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, 65(6):1663–1677, 1994.

[3] A.J. DeCasper and W.P. Fifer. Of Human Bonding: Newborns Prefer Their Mothers' Voices. In Mary Gauvain and Michael Cole, editors, *Readings on the Development of Children*, chapter 8, page 56. Worth Publishers, 2004.

[4] L. Fenson, V. A. Marchman, D. Thal, P.S. Dale, and J. S. Reznick. *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* Paul H. Brookes Publishing Co., 2007.

[5] C.A. Fowler, E.T. Levy, and J.M. Brown. Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language*, 37:24–40, 1997.

[6] L. Gleitman. The structural sources of verb meanings. *Language acquisition*, 1:3–55, 1990.

[7] L. Gleitman and E. Wanner. The state of the state of the art. In *Language acquisition: The state of the art*, pages 3–48. Cambridge University Press, 1982.

[8] L.R. Gleitman, H. Gleitman, B. Landau, and E. Wanner. Where learning begins: initial representations for language learning. In Frederick J. Newmeyer, editor, *Linguistics: The Cambridge Survey*, chapter 6, pages 150–192. Cambridge University Press, 1988.

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

[10] K. Hirsh-Pasek, K. Nelson, G. Deborah, P.W. Jusczyk, K.W. Cassidy, et al. Clauses are perceptual units for young infants. *Cognition*, 26(3):269–286, 1987.

[11] J. Huttenlocher, W. Haight, A. Bryk, M. Seltzer, and T. Lyons. Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27, 1991.

[12] P.W. Jusczyk, K. Hirsch-Pasek, D.G. Kemler Nelson, L.J. Kennedy, et al. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24(2):252–293, 1992.

[13] N.D.G. Kemler, K. Hirsh-Pasek, PW Jusczyk, and KW Cassidy. How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16(1):55–68, 1989.

[14] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118:1038, 2005.

[15] E. Lieven, D. Salomo, and M. Tomasello. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20, 2009.

[16] J. Mehler, P. Jusczyk, G. Lambertz, H. Nilofar, J. Bertoncini, and C. Amiel-Tison. A precursor of language acquisition in young infants. *Cognition*, 29:143–178, 1988.

[17] E.L. Moerk. Processes of language teaching and training in the interactions of mother-child dyads. *Child Development*, 47(4):1064–1078, 1976.

[18] J.L. Morgan. *From simple input to complex grammar*. MIT Press, 1986.

[19] J.L. Morgan, R.P. Meier, and E.L. Newport. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, 19(4):498–550, 1987.

[20] J.L. Morgan and E.L. Newport. The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning & Verbal Behavior. Vol*, 20(1):67–85, 1981.

[21] E.L. Newport, H. Gleitman, and L.R. Gleitman. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow and C. A. Ferguson, editors, *Talking to Children: Language Input and Acquisition*, pages 109–149. Cambridge University Press, 1977.

[22] B.A. Pan, M.L. Rowe, J.D. Singer, and C.E. Snow. Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, 76(4):763–782, 2005.

[23] A. Peters. Language typology, individual differences and the acquisition of grammatical morphemes. In Dan Isaac Slobin, editor, *The cross-linguistic study of language acquisition*, volume 4. Lawrence Earlbaum, 1992.

[24] A.M. Peters. Language segmentation: Operating principles for the perception and analysis of language. In Dan Isaac Slobin, editor, *The crosslinguistic study of language acquisition*, volume 2, pages 1029–1067. Lawrence Erlbaum, 1985.

[25] Ann M. Peters. *The Units of Language Acquisition*. Cambridge University Press, 1983.

[26] J. Piaget. The origins of intelligence in children New York. *International Universities Press, Inc*, 1952.

[27] Brandon C. Roy, Michael C. Frank, and Deb Roy. Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Cognitive Science Conference*, 2009.

[28] Brandon C. Roy and Deb Roy. Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech*, Brighton, England, 2009.

[29] Brandon C. Roy, Soroush Vosoughi, and Deb Roy. Automatic estimation of transcription accuracy and difficulty. In *Proceedings of Interspeech*, Makuhari, Japan, 2010.

[30] Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, and Peter Gorniak. The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference*, pages 2059–2064, Mahwah, NJ, 2006. Lawrence Earlbaum.

[31] Audacity Team. Audacity (version 1.3.4-beta)[computer program]. Retrieved May 5, 2008, http://audacity.sourceforge.net/, 2008.

[32] K. Thorpe and A. Fernald. Knowing what a novel word is not: Two-year-olds 'listen through' ambiguous adjectives in fluent speech. *Cognition*, 100, 2006.

[33] M. Tomasello. *First verbs: A case study of early grammatical development*. Cambridge Univ Pr, 1992.

[34] Soroush Vosoughi, Brandon C. Roy, Michael C. Frank, and Deb Roy. Contributions of prosodic and distributional features of caregivers' speech in early word learning. In *Proceedings of the 32nd Annual Cognitive Science Conference*, 2010.

[35] H. Weide. The CMU Pronunciation Dictionary, release 0.6. Carnegie Mellon University, 1998.

[36] Hung-chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker. Map-reduce-merge: simplified relational data processing on large clusters. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1029–1040, New York, NY, USA, 2007. ACM.

[37] Norimasa Yoshida. Automatic utterance segmentation in spontaneous speech. Unpublished Master's thesis. Massachusetts Institute of Technology, 2002.

[38] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK Book. Cambridge University Engineering Dept, 2006.