Investigating the Role of the Linguistic Environment in Child Language Acquisition using Dense Longitudinal Corpus

by

Soroush Vosoughi

Thesis Proposal for the Degree of
Master of Science in Media Arts and Sciences
at the
Massachusetts Institute of Technology

Fall 2009

Thesis Advisor: _____

Dr. Deb K Roy
Associate Professor of Media Arts and Sciences
MIT Media Laboratory

Thesis Reader: _____

Dr. John Makhoul
Chief Scientist
BBN Technologies

Thesis Reader: _____

Dr. Rochelle Newman
Associate Professor
Department of Hearing & Speech Sciences
Program in Neuroscience & Cognitive Science
University of Maryland

Table of Contents

Executive Summary

What is the relationship between the input that children hear and the words that children acquire? I will investigate the role of the linguistic environment in child language acquisition. This is done primarily by investigating the role of different acoustical and linguistically input variables such as input word frequency and prosody in one child's lexical acquisition using the corpus of Human Speechome Project (HSP) (Roy et al., 2006). For my thesis, I will attempt to create a predictive model of the child's word acquisition based on information encoded in child-available speech .

In order to do this, I need to extract features such as intensity, fundamental frequency, duration, etc from all child-available speech in the HSP dataset. I do this by creating an audio and speech analysis pipeline (Figure 4) that utilizes various speech and audio analysis algorithms such as the Hidden Markov Model Toolkit (Young et al., 2001) and the speech analysis tool praat (Boersma, 2009). Some of the features extracted have certain parameters that need to be optimized. In order to optimize the parameters of a variable, we need to search for parameters that result in a stronger correlation between the saliency of that variable and the age of acquisition of words by the child. In order to achieve this, I have implemented a searcher that is parallelized over 16 cores which uses brute force techniques to check every possible combination of parameters in order to find the optimal parameters for each variable.

After all the relevant variables have been extracted and their parameters optimized, I use linear regression to create a predictive model of the child's word acquisition. Finally, in order to evaluate the model, I look at its predictive power. This is done by doing standard k-fold cross validation of the model. Using this evaluator we can compare the predictive power of different models (with different variables).

Using this system, I can study the relationship between many interesting variables encoded in the child-available speech and the lexical development of child. This will help us better understand the mechanism behind language acquisition.

Though there have been numerous studies of the relationship between child-available speech and the lexical development of children, dating back almost four decades (Snow and Ferguson., 1977; Newport et al., 1988; Furrow et al., 1979; Pinem 1995; Huttenlocher et al.,1991; Goodman et al. 2008), our understanding of the role of child-available speech in language development is limited. This is due to the lack of naturalistic, longitudinal and dense corpora. The HSP corpus, was collected in a way to overcome these limitations which makes my analysis the first of its kind.

Introduction

What role does the linguistic environment play in child language acquisition? Why do children learn e some words earlier than other words? Does the caregivers use of language in the presence of the child have an effect on how child's language development? I will try to answer some of these questions by investigating the role of different acoustical and linguistically input variables such as input word frequency and prosody in one child's lexical acquisition using the corpus of Human Speechome Project (HSP) (Roy et al., 2006).

The HSP corpus is high-density, longitudinal and naturalistic. The corpus consists of high fidelity recordings collected from microphones embedded throughout the home of a family with a young child (Roy et al., 2006). I analyze data collected continuously from ages 9 24 months, including the child's first productive use of language at about 11 months and ending at the child's active use of more than 500 words.

Although the corpus as a whole contains more than 70% of the child's total language input (an estimated 16 million words), I analyze an evenly-sampled 400,000 word portion that has been hand-transcribed using new, semi-automatic methods and for which the speaker has been automatically identified with high confidence (Roy & Roy, 2009). The corpus contains both the child's productions and "child-available speech" which is caregiver speech during which the child was present.

For my thesis, I will attempt to create a predictive model of the child's word acquisition based on information encoded in child-available speech . The information includes variables like input word frequency, intensity, fundamental frequency, duration of phonemes, length of utterances, time of day and many others. I will attempt to predict the age of acquisition of words by the child by analyzing the child-available speech in our corpus. Since through audio alone we can not pinpoint the age of acquisition of words, I instead use the age at which the child first produced a word (called the word birth) as a conservative estimate of age of acquisition. My research will determine what variables in child-available caregiver speech are indeed predictive of the child's language development . This will hopeful help us better understand the nature of child language development.

Proposed Work/ Evaluation and Contribution

Figure 1 shows the five major components that are needed in order to do the child language acquisition study that I have proposed. In this section I will briefly describe each component.

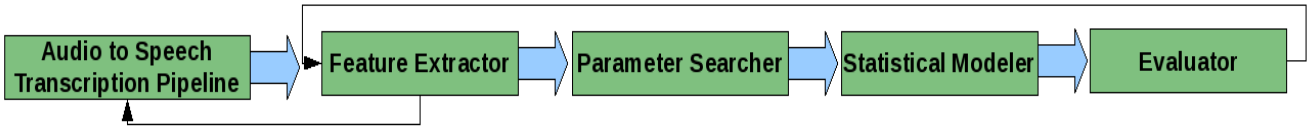

Figure 1. System Architecture

Audio to Transcription Pipeline

The first step is to get from raw audio to transcribed and manageable speech utterances. As mentioned earlier in the proposal, a semi-automated system for achieving this goal has already been developed by Brandon Roy (Roy & Roy, 2009). Figure 2 shows the pipeline that was developed to get from raw audio data to transcribed speech utterances. In order to improve the accuracy and efficiency of the system I will use the transcribed speech to further tune the automated part of the system. I do this by using the transcribed data to train speaker dependent acoustic and language models for each caregiver

and the child using the Hidden Markov Model Toolkit (HTK) (Young et al., 2001). I then feed these models back into the automated speech detector and speaker identifier which can use these models to further tune themselves. The red boxes in Figure 3 show the proposed improvements to the audio/speech pipeline.

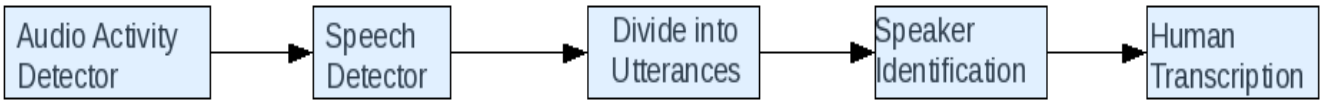

Figure 2. Current Audio to Transcription Pipeline

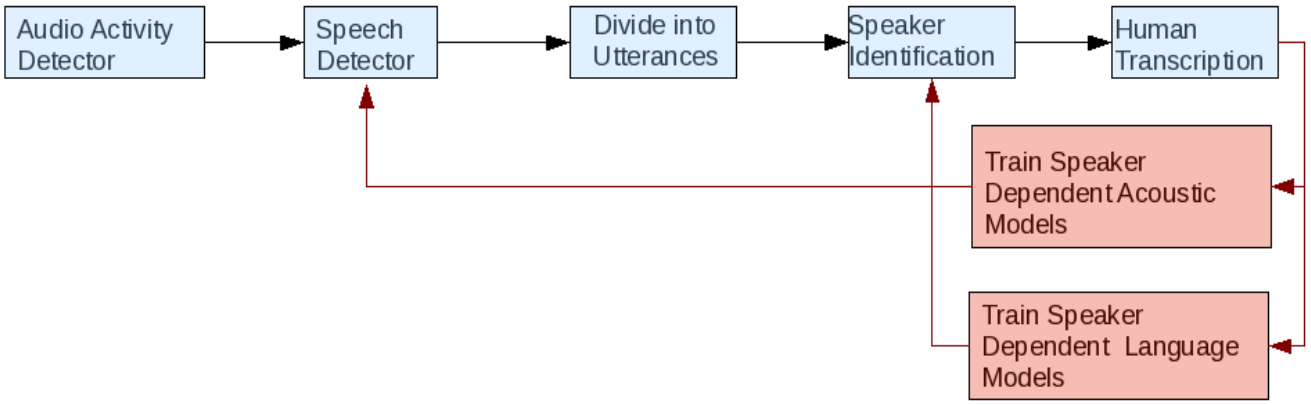

Figure 3. Proposed Improvements to Audio/Transcription Pipeline

Acoustic Feature Extractor

In order to do our analysis, we need to extract features such as intensity, fundamental frequency, duration, etc from all child-available speech. This is done through various speech and audio analysis algorithms. Most notably, the phoneme durations are extracted using the forced-alignment algorithm provided in HTK. Intensity and fundamental frequency are extracted using the speech analysis tool praat (Boersma, 2009). Figure 4 shows the pipeline used to extract these features.

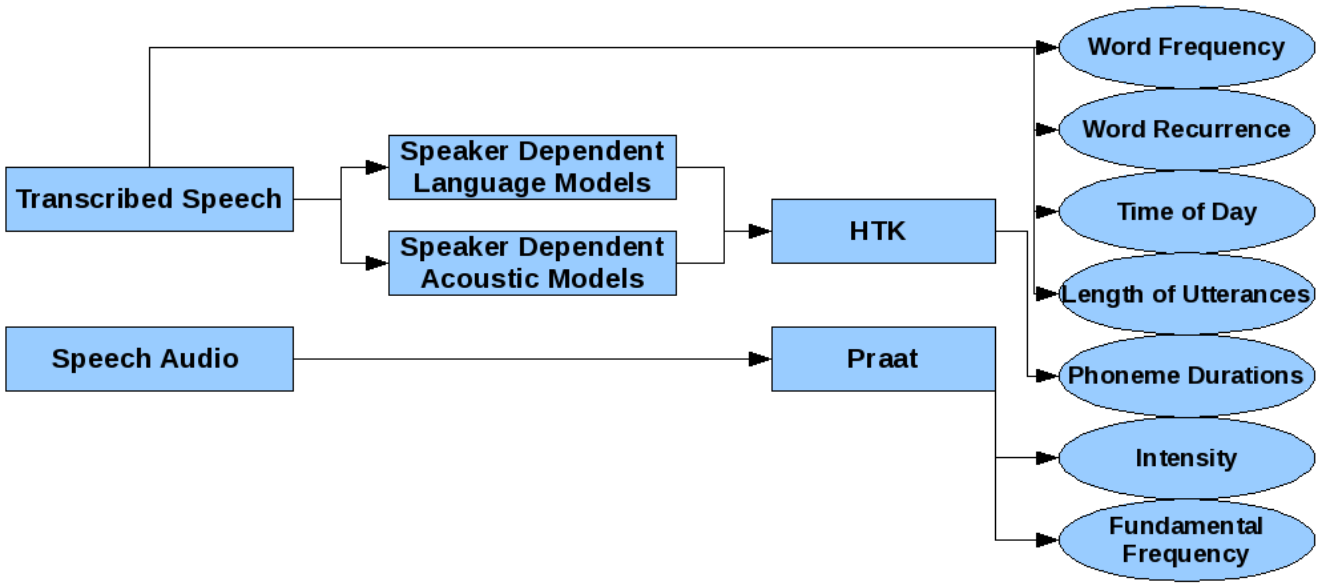

Figure 4. Feature Extraction Pipeline

As you can see from figure 4, currently I am looking at seven different variables in child-available speech. Those are Frequency, Recurrence, Time of Day, Mean Length of Utterances (MLU), Duration, Intensity and Fundamental Frequency. I will be adding many other interesting variables to this list. One variable that might prove to be quite interesting to study is the phonetic distance of the words that the child learns. In other words looking at how phonetic similarity between words would affect the child's learning of new words. I will briefly describe what each of the seven variables listed above encode.

Frequency:
The number of times each word in the child's vocabulary has been uttered by a caregiver before the "birth" of that word, normalized by time.
Recurrence:
The average recurrence of a word (from the child's vocabulary) by caregivers before the birth of that word in a given window size.
Time Of Day:
For each word in the child's vocabulary, we look at the average distance of the time the caregivers uttered the word (before the birth of the word) to sleep time of the child.
Mean Length of Utterances:
For each word in the child's vocabulary, we look at the mean utterance length of caregiver speech (before the birth of the word) containing that word.
Duration:
We use a standardized measure of mean word duration. This is calculated by extracting duration for all vowel tokens, converting these to normalized units for each vowel separately , and then measuring the mean standardized vowel duration for the tokens of a each word in the child's vocabulary.
Intensity:
For each word in the child's vocabulary, we look at the variance of the mean intensity of that word (in all caregiver utterances before the birth of that word) from the mean intensity of the caregiver utterance containing that word.

Fundamental Frequency:
For each word in the child's vocabulary, we look at the variance of the mean F0 of that word (in all caregiver utterances before the birth of that word) from the mean F0 of the caregiver utterance containing that word. We then combine this with the slope of the F0 of that word.

Parameter Searcher

Some of the features extracted in the previous section have certain parameters that need to be optimized (like window size in recurrence described in previous section). In order to optimize the parameters of a variable, we need to search for parameters that result in a stronger correlation between the saliency of that variable and the age of acquisition of words by the child. In other words, we are looking for parameters that increase the predictably power of the variable.

In order to achieve this, I have implemented a parallelized searcher that uses brute force techniques to check every possible combination of parameters and find the optimal parameters for each variable. Figure 5 shows the search process for the optimal window size parameter for the recurrence variable. As you can see in Figure 5, the parameter has tried all possible window sizes from 0-500 and has found the most significant correlation at window size of 51 seconds which produced a correlation of 0.37.

Figure 5. Parameter Searcher Looking For Optimal Window Size For Recurrence

As mentioned before, the searcher checks every possible combination of parameters for every possible combination of variables. This causes the search space to grow exponentially as we add more variables. Just for the seven variables mentioned above, there are hundreds of millions of possible parameter combinations to be checked. In order for the searcher to run in a reasonable amount of time, I have parallelized the searcher using standard Map-Reduce algorithm developed at Google (Dean and Gemawat, 2004). The searcher now runs on four dedicated quad-core machines, giving it a total of 16 cores to use.

Modeler

After all the relevant variables have been extracted and their parameters optimized, I use linear regression to create a predictive model of the child's word acquisition. Figure 6 shows the result of linear regression on the seven variable mentioned before. The red line in Figure 6 below shows the predictive model for the above seven variables. The equation below describes our model.

$$\text{AoA(Word)} = \beta 1 * \text{Frequency} + \beta 2 * \text{Recurrence} + \beta 3 * \text{Time of Day} + \beta 4 * \text{MLU} + \beta 5 * \text{Duration} + \beta 6 * \text{Intensity} + \beta 7 * \text{F0} + \text{Intercept}$$

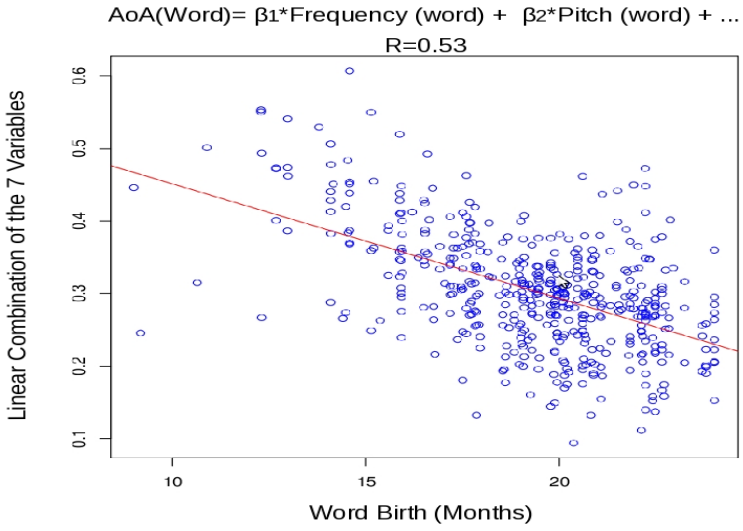

Figure 6. Linear Model of AoA as a Function of Seven Input Variables

Evaluator

Finally, in order to evaluate the model, we look at its predictive power. This is done by doing standard k-fold cross validation of the model. We do this by training our model on all the words in the child's vocabulary except for one. Then we see how accurately our model can predict the birth of the word that was omitted in the training phase. We do this for all words in the child's vocabulary. Figure 7 shows this being done for the word “cake”. The red line represents our model, the blue dot shows the predicted AoA and the green dot is the actual AoA.

Using this evaluator we can compare the predictive power of different models (with different variables) and find the best model. We also compare the predictive accuracy of our model to that of a random model as a baseline.

As shown in Figure 1, the evaluator then feeds back into the feature extractor to tune it into picking better features.

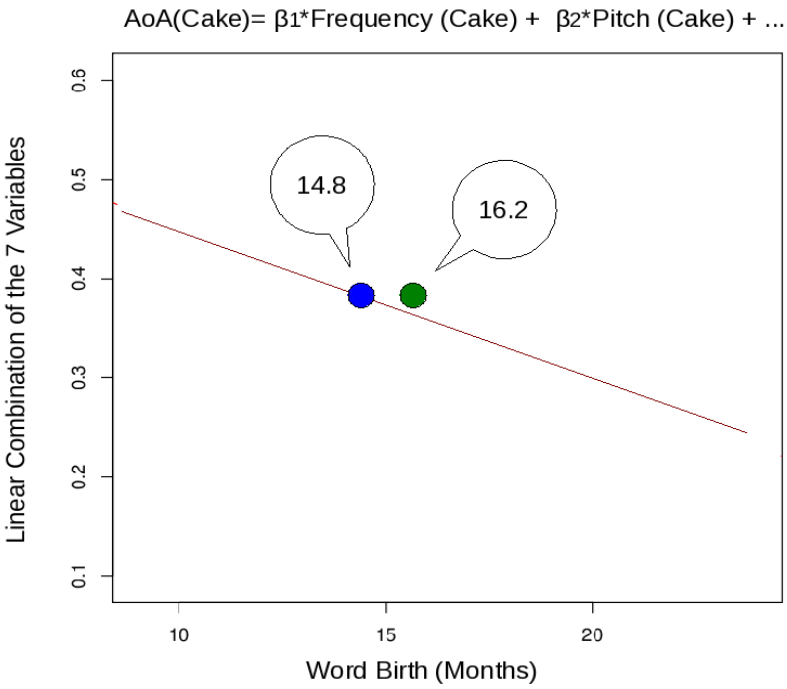

Figure 7. Evaluation of the Model, Using the Word “Cake”

Time Frame

I already have implemented the systems needed for me to do my analysis. I have also already started looking at the correlation between seven variables in child-available speech and the age of acquisition of words by the child. I will spend the majority of my remaining time to study more variables and to do a more thorough evaluation of the system. I will spend months of April and May writing my thesis.

Task	Nov 2009	Dec 2009	Jan 2010	Feb 2010	Mar 2010	Apr 2010	May 2010
Closer look at several new variables.							
Read more related literature							
Thorough Evaluation							
Reader Feedback							
Writing Thesis							

Resources Required

In order to complete my thesis, I need access to the HSP corpus and several fast computers. I already have access to the HSP corpus and the computers. I have already obtained COUHES approval for this study.

References

[1] Boersma, Paul & Weenink, David (2009). Praat: doing phonetics by computer (Version 5.1.02) [Computer program]. Retrieved March 9, 2009, from http://www.praat.org/

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

[3] Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak. (2006). The Human Speechome Project. Proceedings of the 28th Annual Cognitive Science Conference

[4] Dean and Ghemawat (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation,

[5] Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: some simple relationships. Journal of Child Language, 6, 423–442

[6] Gleitman, L. (1990). The structural sources of verb meanings. Language Acquisition, 1, 3-55.

[7] Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language, 35, 515-531.

[8] Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. Developmental Psychology, 27, 236-248.

[9] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Fundamental frequency lends little prominence. Journal of the Acoustical Society of America, 118, 1038-1054.

[10] Newport, E., Gleitman, H., & Gleitman, L.(1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. Talking to children: Language input and acquisition, 109–149.

[11] Pine, J. (1995). Variation in vocabulary development as a function of birth order. Child Development, 66, 272–272.

[12] Young, S., Evermann, G., Kershaw, D., Moore, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2001. The HTK Book. Cambridge University Engineering Dept.

Reader Biography

Dr. Deb Roy directs the Media Lab's Cognitive Machines group and chairs MIT's academic program in Media Arts and Sciences. His research focuses on the interaction of language with physical and social context, which he explores through the construction of robots, video games, and the study of child language acquisition.

Dr. John Makhoul is the Chief Scientist at BBN Technologies. Dr. Makhoul is one of the world's leading experts on speech and signal processing. Dr. Makhoul works on various aspects of speech coding, speech synthesis, speech recognition, speaker identification, artificial neural networks, digital signal processing, optical character recognition, language understanding, speech-to-speech translation, and human-machine interaction using voice. Dr. Makhoul is also an Adjunct Professor at Northeastern University.

Dr. Rochelle Newman is an Associate Professor and the director of graduate studies at University of Maryland's Department of Hearing and Speech Sciences. Dr. Newman is also a faculty member in the Program in Neuroscience and Cognitive, and the Center for Comparative and Evolutionary Biology of Hearing. She is also an affiliate of the Center for Advanced Study of Language at University of Maryland. Her research focuses on speech perception and language acquisition.