

Contributions of Prosodic and Distributional Features of Caregivers' Speech in Early Word Learning

Soroush Vosoughi¹, Brandon C. Roy¹, Michael C. Frank² and Deb Roy¹

{soroush, bcroy, dkroy}@media.mit.edu, The Media Laboratory¹

mcf Frank@mit.edu, Department of Brain and Cognitive Sciences²

Massachusetts Institute of Technology

Abstract

How do characteristics of caregiver speech contribute to a child's early word learning? We explore the relationship between a single child's vocabulary growth and the distributional and prosodic characteristics of the speech he hears using data collected for the Human Speechome Project, an ecologically valid corpus collected from the home of a family with a young child. We measured F0, intensity, phoneme duration, usage frequency, recurrence, and MLU for caregivers' production of each word that the child learned during the period of recording. When all variables are considered, we obtain a model of word acquisition as a function of caregiver input speech. Coefficient estimates in the model help to illuminate which factors are relevant to learning classes of words. In addition, words that deviate from the model's prediction are of interest as they may suggest important social, contextual and other cues relevant to word learning.

Keywords: language acquisition; word learning; corpus data; prosody

Introduction

How does the linguistic environment contribute to children's early word learning? We address this question by making an in-depth study of a single child's vocabulary growth and the relationship of this growth to prosodic and distributional features of the naturally occurring caregiver speech that the child is exposed to. Studying this relationship has the potential to illuminate not only the role of environmental factors in word learning, but also the child's underlying learning mechanisms.

Children's linguistic environments plays a crucial role in determining what they learn, but the precise relationship between what children hear (their input) and what they learn is still unknown. Much of the debate about the role of the linguistic environment has centered around whether the particular properties of child-directed speech (CDS) are useful for the acquisition of syntax. On the one hand, Snow (1986) emphasized the importance of CDS for conveying communicative intent and its consequent importance to development. However, the work of Newport, Gleitman, and Gleitman (1977) challenged the assumption that CDS is a simplified teaching language that facilitates the acquisition of specific syntactic constructions. More recent work has focused on broader patterns of development, documenting a correlation between grammatical and lexical developmental trajectories (Bates & Goodman, 1999).

Stronger evidence for the contributions of CDS to language development have been found in the realm of lexical acquisition. For example, Huttenlocher, Haight, Bryk, Seltzer, and

Lyons (1991) found a positive correlation between the quantity of CDS and a child's vocabulary size and rate of growth. Increased frequency of use of particular words in CDS has also been tied to earlier acquisition of those words by the child (Huttenlocher et al., 1991; Goodman, Dale, & Li, 2008; Roy, Frank, & Roy, 2009). Frequency is not the only factor that affects acquisition, however. The production of a word in isolation is also a consistent predictor of lexical development (Brent & Siskind, 2001). Finally, prosodic factors in caregiver speech also likely play a role in acquisition: Echols and Newport (1992) found that children were much more likely to produce and recognize syllables that were stressed in caregivers' speech.

While previous studies of the relationship between CDS and children's vocabulary acquisition have largely focused on examining a small section of the input to a range of children, here we take a different approach. We make a very detailed study of this relationship in a very large, dense, longitudinal dataset collected in an ecologically valid setting. This dataset was collected as part of the Human Speechome Project (Roy et al., 2006). At present, the Speechome Corpus consists of time aligned orthographic transcripts as well as a complete audio and video record of all data collected. Therefore, our analysis is not limited to factors like frequency (which can be computed from transcripts alone): instead we are able to include additional prosodic variables that can only be computed from aligned audio and transcripts.

Our goal in this current analysis is to predict the child's age of acquisition (AoA) for individual words on the basis of information from CDS. AoA is usually categorized as the age of *receptive* and *productive* acquisition. Receptive acquisition is typically determined by the caregiver via diary studies or checklists, and is consequently relatively difficult to assess with high accuracy for a large sample of words. Age of productive acquisition is more easily measured from transcripts, although there are complications here as well, since early productive word forms often differ from the corresponding adult word form. However, we are able to overcome this limitation to a greater extent than previous studies, because of the density of our data and the accessibility of caregivers for help in the transcription process.

The plan of our paper is as follows. We begin with an overview of the Human Speechome Project. We then review the regression framework we used for the prediction of vocabulary acquisition and describe in detail the predictors we included in this framework. We report both simple correlations

between individual predictors and age of word acquisition as well as the results of a series of regression models. We end by considering the implications of our work for future research in language acquisition.

The Human Speechome Project

The Human Speechome Project (HSP) (Roy et al., 2006) 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 house of one author's (DR) family was outfitted with fourteen microphones and eleven omnidirectional 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 fisheye lens video cameras that provide a bird's-eye view of people, objects, and activity throughout the home.

The Speechome project captures one child's development in tremendous depth. While this aspect of the project limits conclusions about general aspects of language development, the dense sampling strategy affords many advantages over other corpora (eg. (Lieven, Salomo, & Tomasello, 2009)). First, the Speechome 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).

The current study builds on our first analysis of the Speechome data (Roy et al., 2009). In that study, we focused on the child's 9-24 month age range and explored several aspects of word learning, examining variables such as the child's vocabulary growth, increase in mean length of utterance (MLU) as well as properties of caregiver speech such as caregiver MLU over time. Due to the high density of data, with several days per week fully transcribed over the course of this 9-24 month period, a surprising picture emerged of the tuned relationship between the child's development and caregiver speech. Congruent with other reports, we found that words used more frequently in caregiver speech tend to be learned earlier by the child, with a much stronger effect when words are grouped by class (Huttenlocher et al., 1991; Goodman et al., 2008).

Methods

The Speechome Audio 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. We have developed a semi-automated speech transcription system called BlitzScribe that facilitates fast and accurate speech transcription (Roy & Roy, 2009). 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 smoother transcription process. We have obtained 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 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).

Our current study 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 72 fully transcribed days containing an average of 23,055 words per day of combined CAS and child speech, totaling 1.66 million words. We estimate that the fully transcribed 9-24 month corpus will contain 12 million words. Our 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 1.66 million words of transcripts and only use a subset of the transcripts for the current analyses. First, roughly 700,000 words belong to utterances marked by human transcribers as containing more than one speaker. In other words, about 40% of pause separated spoken utterances contain abutting or overlapping speech of two or more people, reflecting the realities of "speech in the wild." Since our objective here is to examine interaction of CAS and child speech, and since we cannot currently 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 399,141 word corpus used for all analyses in this paper.

Outcome and Predictor Variables

The goal of our study was to use measurements of the prosodic and distributional characteristics of CAS to predict

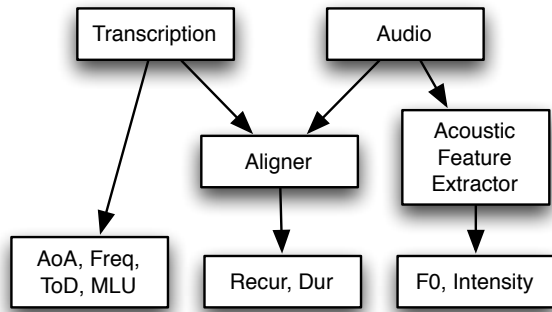


Figure 1: Schematic of the processing pipeline for outcome and predictor variables.

AoA for the child’s early vocabulary. We use linear regression to provide a computational framework for this goal. We therefore used age of acquisition as our outcome variable and extracted six predictor variables to quantify aspects of CAS. Figure 1 shows the pipeline used to extract these predictor variables from our speech and transcription files. Below we give our operational definition for age of acquisition and for each of the six predictor variables we used in our analysis. All variables are computed using the sample up to the AoA for a particular word.

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.

Frequency Frequency measures 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.

Recurrence Distinct from frequency, recurrence measures the repetition of a particular word in caregiver speech within a short window of time. The window size parameter was set by searching all possible window sizes from 1 to 600 seconds. 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 (51 seconds).

MLU The MLU predictor measures the mean utterance length of caregiver speech containing a particular word. In order to be consistent with the direction of correlation for

other variables (a negative correlation with the AoA) we use $1/\text{MLU}$ as the predictor.

Duration The duration predictor is a standardized measure of word duration for each word. We first extracted duration for all vowel tokens in the corpus. We next converted these to normalized units for each vowel separately (via z-score), and then measured the mean standardized vowel 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 vowel that occurred in tokens of “dog” was often long relative to comparable vowel sounds that appeared in other words. We grouped similar vowels by converting transcripts to phonemes via the CMU pronunciation dictionary.

Fundamental frequency 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 (Boersma & Weenink, 2009). We then calculated the change in F0 as a sum of two terms shown in the equation below. The first term captures the change in F0 for the word relative to the utterance in which it’s embedded. $\overline{F0}_w$ is the mean F0 value of the word, and $\overline{F0}_{utt}$ is the mean F0 of the whole utterance. The second term captures the maximum change in F0 within the word. t_{max} and t_{min} are the times at which the max and min F0 values occur within the word. α_0 and α_1 are constants set using the same optimization technique described in the recurrence section.

$$\alpha_0 * |\overline{F0}_w - \overline{F0}_{utt}| + \alpha_1 * \left| \frac{\max(F0_w) - \min(F0_w)}{t_{max} - t_{min}} \right|$$

Intensity Relative word intensity was calculated in the same manner as F0 using the intensity contour in place of the F0 contour. The intensity contour was extracted using the PRAAT system.

Results

Correlation analysis

Correlations between AoA and the six variables we coded in caregiver speech are shown in Figure 2. All correlations 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.

Replicating results in Roy et al. (2009), the correlation with frequency was -.23. This figure is slightly lower than the -.29 reported in the earlier paper. There are two differences in analysis that account for the different result. First, a small subset of words were excluded from this analysis due to data sparsity. Second, frequency data are estimated only up to the time the child first produces the word. This second difference leads to a potentially interesting conclusion. If the distribution of word frequencies is stationary with respect to time, then correlations should go up as more data are included for

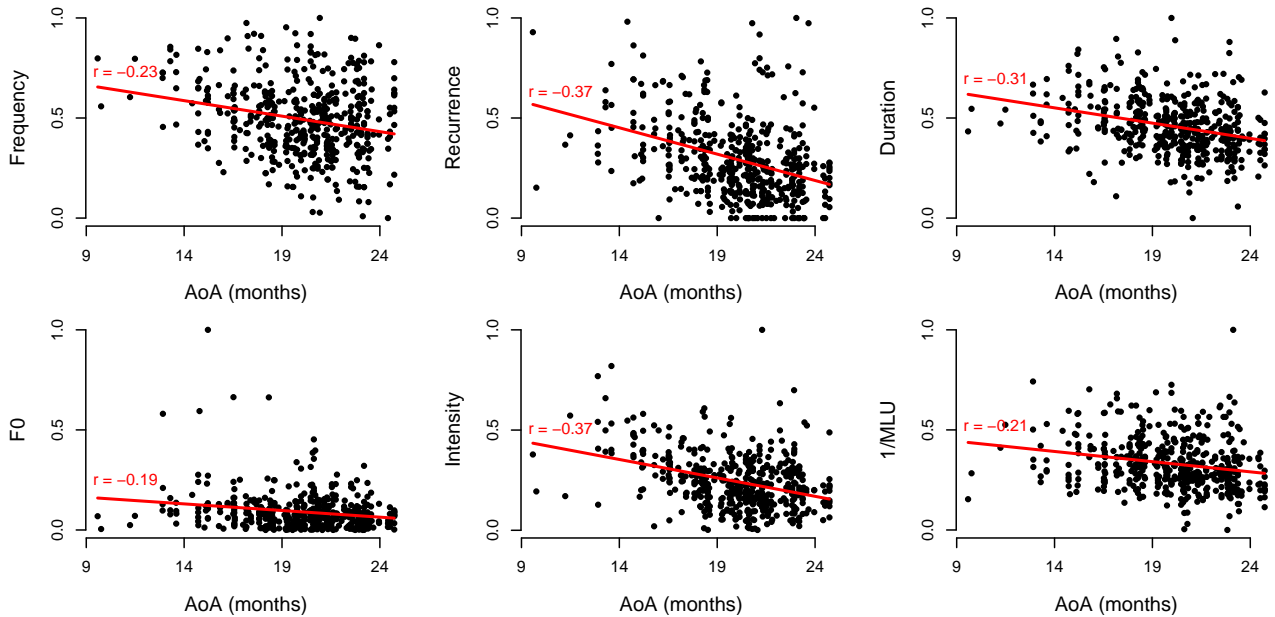


Figure 2: Each subplot shows the univariate correlation between AoA and a particular predictor. Each point is a single word, while lines show best linear fit.

Table 1: Correlation coefficients (Pearson’s r) between all predictor variables. Note: $^{\prime} = p < .1$, $* = p < .05$, and $** = p < .001$.

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

each word. In contrast, if caregivers tune the frequency distribution of words to an estimate of the child’s knowledge, correlations should go down as more data are included. Because we observed (slightly) larger correlations with frequency for the earlier dataset, this provides some evidence against caregiver tuning of word frequencies.

Correlations between predictor values are shown in Table 1. 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. On the other hand, correlations between prosodic variables like frequency and intensity or duration and inverse MLU are less clear. For example, perhaps words are more likely to have longer duration vowels when they are being accented in a shorter sentence.

Regression analysis

We next constructed a regression model which attempted to predict AoA as a function of a linear combination of predic-

tor 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 (Fenson, Marchman, Thal, Dale, & Reznick, 2007) 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 ($F(4) = 107.37$, $p < .001$).

Coefficient estimates for each predictor are shown in Figure 3. All predictors were significant at the level of $p < .05$. The full model had $r^2 = .32$, 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,” “hmm,” and “sarah”). 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.

Assessing model fit

Residuals from the basic linear model were normally distributed. Figure 4 shows the relation between predicted age of acquisition (via the full predictive model including part of

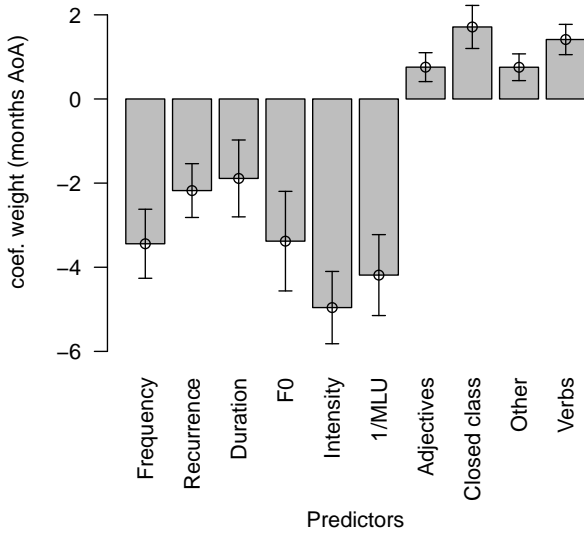


Figure 3: Coefficient estimates for the full linear model including all six predictors (and part of speech as a separate categorical predictor). Nouns are taken as the base level for part of speech and thus no coefficient is fit for them. Error bars show coefficient standard errors. For reasons of scale, intercept is not shown.

speech) and the age of acquisition of words by the child. One useful aspect of plotting the data in this way 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 “nannynname” (proper names have been replaced for privacy reasons) are learned far earlier than predicted by the model (above the line of best 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. Future work will use these errors of prediction as a starting point for understanding contextual factors influencing word learning.

Interactions and more complex models

Our first linear model had 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 2 shows coefficient estimates for five linear mod-

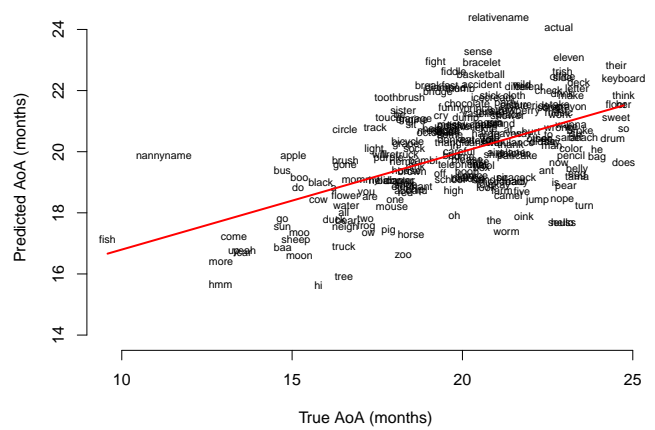


Figure 4: Predicted AoA vs. true AoA. To avoid overplotting, only half of the 461 words are shown. The red line shows the line of best fit.

Table 2: Coefficient estimates for linear models including data from adjectives, nouns, closed-class words, verbs, and all data. Note: ‘ = $p < .1$, * = $p < .05$, and ** = $p < .001$.

	Adj.	Closed	Nouns	Verbs	All
Icept	27.66**	25.03**	25.00**	25.93**	25.57**
Freq	0.38	6.73	-5.84**	-0.89	-1.53*
Recur	-2.36	-12.02*	-1.53’	-7.47**	-2.85**
Dur	-5.22*	1.81	0.09	-2.74	-2.66*
F0	-7.43	-6.42	-2.28’	0.54	-3.42*
Int.	-8.60*	-12.16	-4.66**	-1.56	-4.78**
1/MLU	-5.70*	-9.37	-3.71*	-5.26	-3.89**

els, 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 (Thorpe & Fernald, 2006). 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 (Gleitman, 1990).

We next constructed a model that included every pairwise interaction between each of the six predictors and between the predictors and POS. We then used stepwise regression to remove predictors that did not increase model fit. Stepwise regression prunes predictors using AIC, a measure which balances increases in likelihood with complexity. This model increased r^2 to .44, and added a large number of interaction

terms. We report only the general outlines of results in this model as they confirm intuitions from other analyses.

While frequency had an overall *positive* coefficient value in this model, all four interactions were negative, indicating that there was considerable shared information between frequency and other predictors. Recurrence and intensity also interacted significantly, suggesting that when words were spoken repeatedly with high intensity (possibly because they were a topic of discourse over a period of time) they were acquired at earlier ages. Finally, both duration and intensity interacted with POS, with significant coefficients for closed-class words. As seen in Table 2, longer closed-class words are acquired slightly later (probably because longer closed-class words are less frequent). In addition, higher intensity closed-class words are acquired considerably earlier, probably because one major challenge in function word acquisition is understanding their prosodic structure (Demuth & McCullough, 2008).

Discussion and Future Work

Our study quantified six variables describing the prosodic and distributional characteristics of words in child-available caregiver speech: frequency, recurrence, mean length of utterance, duration, fundamental frequency, and intensity. We found that each of these variables helped to predict the age at which the child acquired words. There were considerable differences in the predictive power of each variable across different parts of speech, however. For example, frequency and intensity mattered most for nouns, while recurrence in a small window of time seemed to matter more for verbs and closed-class words. These results complement previous smaller-scale, cross-sectional investigations and provide a variety of new directions for potential experimental manipulations.

Our current model only takes 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 current investigation has been limited to the child's lexical development; our plan is that future work will extend the current analysis to grammatical development.

Finally, the analysis and findings presented in this paper assume a linear input-output model between child and caregivers: the caregivers produce input to the child, who then learns. 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 our own previous findings (Roy et al., 2009). Our ongoing work continues to investigate the mutual influences between caregivers and child and to measure the degree of adaptation in this dynamic social system.

References

- Bates, E., & Goodman, J. (1999). On the Emergence of Grammar from the Lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Lawrence Erlbaum Associates.
- Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1.01)*. <http://www.praat.org/>.
- Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 33–44.
- Demuth, K., & McCullough, E. (2008). The prosodic (re) organization of children's early english articles. *Journal of Child Language*, *36*.
- Echols, C., & Newport, E. (1992). The role of stress and position in determining first words. *Language acquisition*, *2*.
- Fenson, L., Marchman, V. A., Thal, D., Dale, P., & Reznick, J. S. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual*. Paul H. Brookes Publishing Co.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*, 3–55.
- Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531.
- Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*.
- Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*.
- Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition* (pp. 109–149). Cambridge University Press.
- Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Cognitive Science Conference*.
- Roy, B. C., & Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech*. Brighton, England.
- Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 2059–2064). Mahwah, NJ: Lawrence Erlbaum.
- Snow, C. E. (1986). Conversations with children. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development* (pp. 69–89). Cambridge, UK: Cambridge University Press.
- Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds 'listen through' ambiguous adjectives in fluent speech. *Cognition*, *100*.