Exploring Word Learning in a High-Density Longitudinal Corpus

Brandon C. Roy (bcroy@media.mit.edu)
The Media Laboratory
Massachusetts Institute of Technology

Michael C. Frank (mcfrank@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Deb Roy (dkroy@media.mit.edu)
The Media Laboratory
Massachusetts Institute of Technology

Abstract

What is the role of the linguistic environment in children's early word learning? Here we provide a preliminary analysis of one child's linguistic development, using a portion of the high-density longitudinal data collected for the Human Speechome Project. We focus particularly on the development of the child's productive vocabulary from the age of 9 to 24 months and the relationship between the child's language development and the caregivers' speech. We find significant correlations between input frequencies and age of acquisition for individual words. In addition, caregivers' utterance length, type-token ratio, and proportion of single-word utterances all show significant temporal relationships with the child's development, suggesting that caregivers "tune" their utterances to the linguistic ability of the child.

Keywords: Language acquisition; word learning; corpus data

Language Development and the Environment

What is the role of the linguistic environment in a child's acquisition of language? In attempts to understand the nature of the mechanisms underlying language acquisition, input-uptake correlations have the potential to provide deep insight. If particular aspects of children's input are predictive of their later language development, such findings could powerfully illuminate the nature of children's language learning strategies and mechanisms and the relationship of linguistic knowledge and experience.

Systematic studies of child-directed speech (CDS) dating back to the late 1960's have established that CDS has special characteristics including shorter utterance lengths, exaggerated prosody, high redundancy, and referential content tied to immediate context (Snow & Ferguson, 1977). Initial investigations of the facilitative role of CDS focused primarily on development of syntax. Early findings were contradictory, however, and the overall picture remains mixed (Newport et al., 1977; Furrow et al., 1979; Pine, 1995). More recent studies of the role of the environment on lexical development proved to be clearer. For example, studies have shown that the total amount of CDS predicts children's vocabulary size and rate of growth (Huttenlocher et al., 1991; Hart & Risley, 1995) and the frequency of specific words within CDS predicts the age of acquisition of those words (Huttenlocher et al., 1991; Goodman et al., 2008).

Despite the quantity of work in this area, however, our overall understanding of the role of the environment in language development remains limited by the lack of appropriate observational data. Historically, most longitudinal studies of language development have relied on observations from just two or a few points in time, leading to difficulties in constructing a complete picture of the continuous developmental process. Driven by new technologies, the methodological landscape is now changing. Higher density longitudinal studies are emerging that provide valuable new perspectives on long-standing questions. For example, by analyzing 90-120 minute audio recordings of children's home environments every two weeks from 9-15 months of age, Brent & Siskind (2001) shed new light on the role of words spoken in isolation by showing that their presence in CDS predicted age of acquisition of those words. More recently, Lieven et al. (in press) recorded 28-30 hours of audio over a 6-week period in the homes of four toddlers yielding 100,000+ word corpora of CDS and speech by children from each home. These data were used to trace the relationship between a child's utterances over time in support of a constructivist theory of grammar development.

Motivated by the goal of obtaining a more complete and naturalistic longitudinal record of child development—and establishing new tools and methods for replicating such efforts in the future—the Human Speechome Project (HSP) was launched with the aim of recording the first two to three years of one child's development at home in rich detail (Roy et al., 2006). This paper provides an overview of the HSP project and corpus, and the human-machine collaborative process for audio analysis. We then present an initial analysis on a subset of the audio portion of this corpus focusing on CDS and lexical development.

The Human Speechome Project

The goal of HSP is 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 the family of one of the authors (DR) with a newborn was outfitted with fourteen microphones and eleven omnidirectional cameras. 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.

Recordings were made from birth to the child's third birthday with the highest density of recordings focused on the first

two years. Given our current interest in early word learning, the analyses below are based on the audio from 9-24 months.

As with previous longitudinal case studies (Piaget, 1952; Tomasello, 1992), conclusions about the general nature of language development that may be drawn from analysis of the Speechome corpus are inherently limited since the data charts only one child's development. However the corpus differs from previous case studies in important respects. In contrast to diary studies which are necessarily theory-laden (since diarists cannot record everything, they must rely on theoretical biases to decide what is noteworthy at the time of observation), the Speechome corpus may be re-analyzed multiple times guided by different theoretical perspectives. In addition, the presence of high-resolution video provides opportunities to study the role of various aspects of non-linguistic context from joint attention to routine activities and beyond.

Human-Machine Data Annotation

Our study of early word learning depends on high accuracy orthographic (word-level) transcription and speaker identification of both child and caregiver speech. In order to circumscribe this task, we limited speaker identification to the child and the three primary caregivers (mother, father, and full-time nanny). 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). However, the size of the corpus still presents a daunting challenge. Typical speech transcription tasks take an order of magnitude longer than the actual single-track audio duration, and with several thousand hours of 14 track audio this approach would be too time-consuming and too costly.

Speech detection, transcription, and speaker ID

To this end we have developed BlitzScribe (Roy, 2007), a system that enables accurate, rapid speech transcription. With this system, automatic audio processing algorithms are used to robustly detect speech in the audio and split speech into short, easy to transcribe segments. Sequences of speech segments are loaded into a specially designed transcription interface that enables a human transcriber to simply listen and type, obviating the need for manually finding and segmenting speech or explicitly controlling audio playback. As a result, playback stays synchronized to the transcriber's speed of transcription. We have found our system to have comparable accuracy (Roy & Roy, under review) to other popular transcription tools yet is approximately five times faster for naturalistic audio recordings. Transcribers using this system can obtain average transcription times of less than twice the audio duration. In addition to BlitzScribe, we have also developed speaker identification algorithms to automatically label speech segments with the speaker, along with associated confidence values. Speaker annotation tools allow a human to review low confidence segments and make corrections as necessary. Using this suite of tools, we expect that a small team of several annotators, working at 40 hours per week, will be able to transcribe the corpus in less than a year.

The Speechome Audio Corpus

In the 9-24 month age range, the Speechome audio corpus contains 4260 hours of 14-track audio of which an estimated 1150 hours contain speech. Of the 488 days that constitute this age range, recordings were made on 444 of the days with a mean of 9.6 hours recorded per day. At the time of this writing, 72 of these days have been fully transcribed with a mean of 23,055 words per day of combined CAS and child speech for a total of 1.66 million words. These 72 days of transcribed speech are evenly distributed over the 16 month range. Based on these statistics, the corpus contains an estimated 10.2 million words of CAS and child speech in total over the 9-24 month range. 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, we removed these utterances since we cannot currently distinguish the source of speech. 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 the bottom ˜50% of utterances. Third, about 15% of the remaining utterances were deemed by human transcribers to be of insufficient clarity to reliably transcribe. After removing those utterances, we obtained the 399,141 word corpus used for all analyses in this paper.

The Child's Productive Lexicon

While the current level of resolution does not allow us to pinpoint exactly how much speech the child produced or the absolute first time the child produced a token of a word, the current density (about a day in every week is transcribed) does allow us to estimate for the first time a variety of different continuous measures. We begin by presenting descriptive measures of the child's language development in this section, and move on to analyses in the next section that attempt to capture coordination between the caregivers and the child.

The child as conversational participant

The child's language development is visible even at a very coarse grain of analysis. Figure 1 shows the proportion of all utterances attributed to each speaker with high confidence by the speaker ID system. The child goes from producing essentially no transcribable speech at 9 months to producing

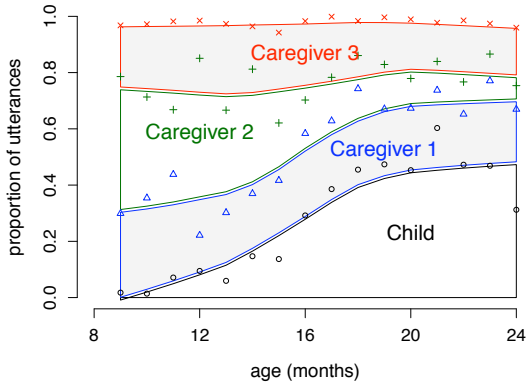

Figure 1: Proportion of all utterances produced by each of the speakers in our study. Points are stacked (cumulative for each month), and lines show the result of a lowess (local linear) smoothing function.

nearly as many utterances as the three other caregivers combined. Particularly striking is the increase that happens between 15 and 17 months, a period during which the number of utterances produced by the child nearly doubles.

Word births

Our next analysis examined the child's lexical development on a word-by-word basis. We did this by defining a *word birth*: the first time in our transcripts that a word entered the child's productive vocabulary. We used this technique to extract a set of candidate word births automatically from corpus data. We then analyzed this list by hand, removing morphological variants of already-born words (e.g., gerunds and plurals) and categorizing words into the categories listed on the MacArthur-Bates Communicative Development Inventory (CDI) (Fenson et al., 1994). We found 517 word births in the corpus by these criteria, of which 265 appeared in the CDI checklist.

The distribution of word births by age was smooth but highly non-linear (Figure 2, top). Word births as measured by our technique increased until 20 months, at which point they decreased quickly.[1] In addition, the category structure of the child's vocabulary shifts rapidly over the course of the period between 9-20 months (Figure 2, bottom). At 20 months, corresponding to the peak of word births observed in the top panel, the composition of the lexicon appears to stabilize.

Word frequency analyses

Relations between input and output have often been investigated in the domain of frequency, both in the general frequency of CDS (Hart & Risley, 1995) and in connections between the specific frequencies of words in CDS and the age at which they are acquired (Huttenlocher et al., 1991; Goodman et al., 2008). Our goal was to replicate this second set of analyses in the Speechome corpus, evaluating correlations between the frequency of words in CDS and their age of acquisition according to the word-birth analysis reported above.

[1] We believe this pattern of increase and decrease is the result of several interacting factors. First, as the child's vocabulary grows larger, the probability of observing a new word sampled from that vocabulary grows smaller. Second, as the child learns the most frequent words in the language, newer words will for the most part be less frequent words and hence will be less likely to be detected even in dense sampling.

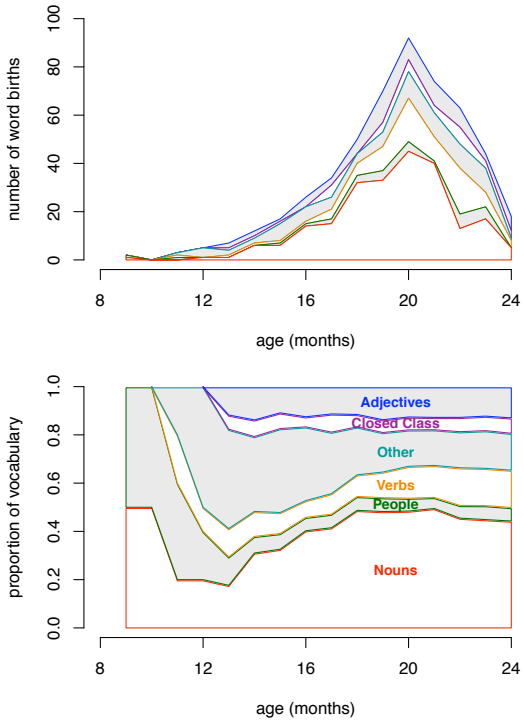

Figure 2: (top) Number of words from each category that were born (used for the first time) at each age. (bottom) Proportion of the child's vocabulary in different word categories at each age. In both plots, regions are stacked so that the total height is a sum across regions. Colors denote categories from Goodman et al. (2008).

To perform this analysis, we regressed the word birth date for each word in the child's productive vocabulary against the total, non-normalized log frequency of that word in the speech of the three caregivers. The results of this analysis are shown in Figure 3, left. We found a significant correlation between frequency and age of acquisition across all words ($r = -.29$, $p < .0001$). While Goodman et al. (2008) found a positive correlation across all words, their CDS corpus contains speech directed at older children and may contain a higher proportion of closed class words. This could be one factor leading to different results. However, when we investigated individual groups of words (as shown in the CDI data), we found heterogeneity among them (examples are shown in Figure 3, middle and right panels). In general, replicating the results of Goodman et al. (2008), noun categories were most highly correlated with frequency, while closed class words were least correlated. It should be noted that only words acquired by 24 months were used in the regression, potentially

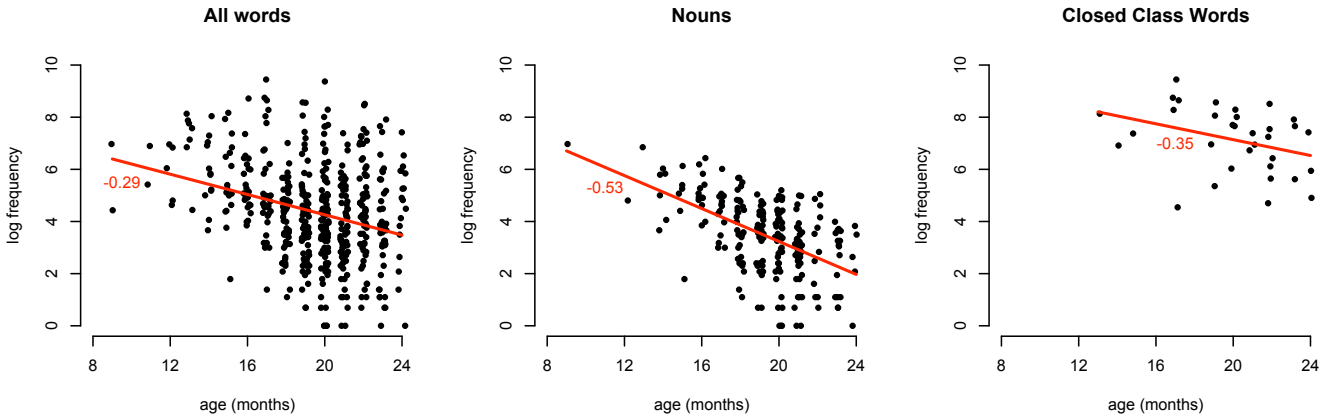

Figure 3: Words plotted by the age they were first produced and their log frequency, along with the best linear fit and *r*-value. Ages are jittered slightly on the x axis to avoid overplotting.

overestimating or underestimating the correlation. We were also interested in whether these correlations held when only the caregiver word counts prior to a word birth were considered. As Goodman et al. (2008) point out, it may be that an earlier word birth leads to higher overall usage of that word by caregivers. However, if caregiver word frequency is calculated only up to the word birth date, the correlations decrease somewhat but the effect remains. Thus, within speech to a single child, we found a strong relationship between input frequency and age of acquisition, although this relationship was presumably mediated by a number of other factors, including syntactic category.

Tuning of caregivers' utterances

Claims about the facilitative role of caregivers' speech have typically hinged on the concept of "tuning": whether caregivers change the form of their utterances in order to accommodate the linguistic knowledge of the child. In our next analyses, we studied the form of caregivers' utterances. Our goal was to test two versions of the tuning hypothesis. First, we were interested in characterizing what we will call "coarse tuning": adjustment of caregivers' speech to the general linguistic competence of the child. Coarse lexical tuning predicts that a caregiver will generally use shorter utterances that are less lexically diverse when talking to a child whose vocabulary and word-combination abilities are limited.

The second version of the tuning hypothesis we call "fine lexical tuning": adjustment of caregivers' speech at the level of individual lexical items. (The term"fine tuning" has previously been used in the literature simply as a catchall term for tuning phenomena, but here we distinguish different levels of the tuning hypothesis). Fine lexical tuning predicts that caregivers will not only adjust their speech to the general level of the child, they will also adjust the complexity of individual utterances on the basis of the familiarity of the child with their lexical content.

Coarse tuning

In our first set of analyses, we computed three descriptive measures of caregivers' speech for each month in our sample. Our hypothesis was that if caregivers were engaging in coarse tuning, we would see temporal changes in the characteristics of their speech coordinated with the linguistic development of the child. The first measure was the mean length of caregivers' utterances (MLU) (Moerk, 1976).[2] Our second measure was the proportion of single-word utterances in caregivers' speech. This measure was directly related to MLU, but since single word utterances had previously been identified as a significant source of information in early word learning (Brent & Siskind, 2001), we computed this measure separately. The final measure was the type-token ratio (TTR) in caregivers' speech. TTR is a measure which captures the lexical diversity of a set of contexts. A high TTR for a word suggests that a word appeared in diverse contexts which shared few words with one another; in contrast, a low TTR suggests that the word appeared in a smaller range of contexts or contexts which shared many words with one another.

Results of these analyses are shown in Figure 4. We found significant correlations between the child's MLU trajectory and the MLU trajectories of all three of his caregivers (Pearson's *r* was .55, .57, and .62, with *p*-values of .03, .02, and .01, for C1-C3 respectively). The child's utterances were exclusively one word long in the period from 9 to 15 months. In his 16th month, his MLU began to increase and rose rapidly thereafter (to a level of ˜2.6 in the month of his second birthday). Caregivers' speech appeared to decrease in MLU until about the time when the child began combining words around 16 months; at this point their MLUs all began increasing. Although more detailed analyses are necessary, this pattern may correspond to a shift in the proportion of utterances for which

[2]We report MLU based on the number of orthographic words—as opposed to morphemes—in each utterance segmented by our automatic system. Thus the MLU figures reported may not be directly comparable to estimates of MLU computed via other methods.

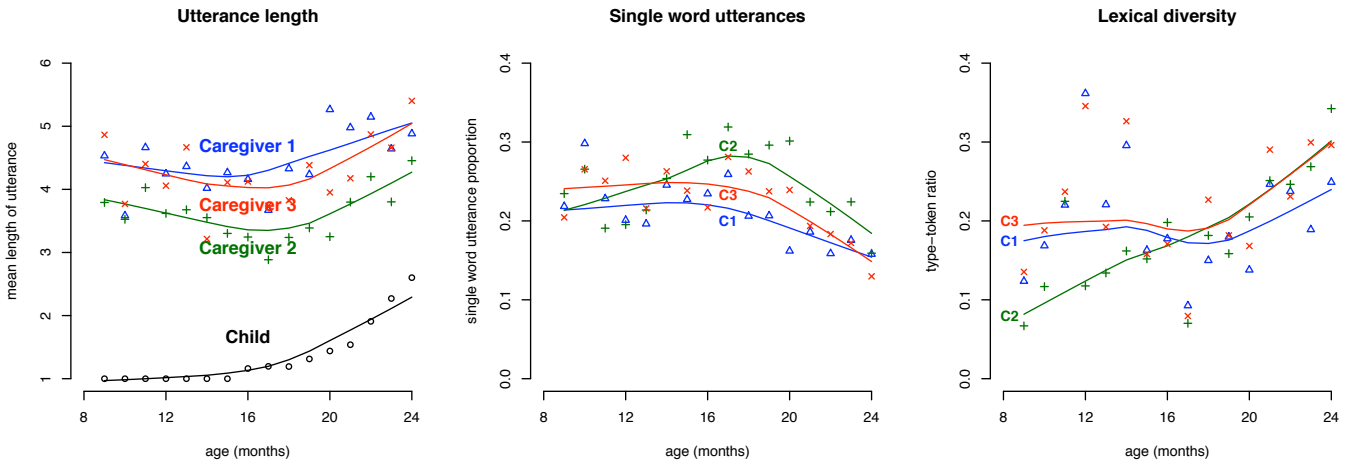

Figure 4: Three measures of utterance complexity for each caregiver, plotted by the age of the child. (left) The mean length of utterances for the child and the three caregivers. The line of best fit for a lowess (local linear fit) smoothing function is shown for each set of points. (middle) The proportion of single word utterances in caregivers' speech. (right) The type-token ratio of caregivers' speech.

the child's understanding was useful or necessary.

The other two measures we computed (single-word utterance proportion and TTR) have not typically been used as measures of the complexity of children's speech, so we examined only in the caregiver data and looked for change over the period of interest. Our measure of single-word utterance prevalence varied across caregivers but showed significant temporal structure. Two caregivers' single-word utterance proportions showed a significant decrease across the period of recording in a linear regression against age (for C1, $p = .002$ and for C2, $p = .01$) while C3 showed no linear trend ($p = .83$) but seemed to show some quadratic trend with single-word utterance proportion peaking around the same time-period as the dip in MLU. In contrast, when we examined the TTR data, C2 exhibited a very strong increasing linear trend with age ($p < .001$) and no trend (or possibly a slight uptick towards the end of recording) for C1 and C3. Although it must remain speculative with such a small sample of caregivers, there is the possibility that the different caregivers mitigated the complexity of their utterances with different strategies, with C1 and C3 using more single word utterances and C2 using highly restricted contexts.

To summarize: we saw significant temporal structure in both the length and lexical diversity of caregivers' utterances, indicating a relationship between the linguistic abilities of the child and the speech of the caregivers. These data add to the literature supporting coarse tuning and suggesting that caregivers adapt the length and lexical complexity of their utterances to what they think their child can understand.

Fine lexical tuning

Our final analyses tested the hypothesis of fine lexical tuning—that caregivers not only tailor their utterances to the general linguistic ability of the child, but that they may also adjust the complexity of the contexts in which they present individual words depending on the child's understanding of those words.

To carry out this analysis, we re-examined the MLU data for each caregiver. For each word in the child's productive vocabulary, we extracted the MLU for each month for the utterances containing that word. This resulted in an MLU time-series for each caregiver for each word; we re-centered these time-series so that they were aligned by the word birth (when the child first produced the word). We then calculated the change in MLU for each word relative to the month in which the child began producing it, and averaged across words.[3] Intuitively, this analysis allow us to look at whether there is a consistent change in caregiver MLU before and after the child knows how to use a word. The results are shown for each caregiver in Figure 5. We found a systematic decrease in the MLU of words immediately prior to the child's first production of that word, without a corresponding rise afterwards. Both the directionality and the magnitude of these changes was different than those observed in the earlier analysis of coordination in MLU and suggest that there is likely some level of fine lexical tuning in caregivers' speech. These changes additionally suggest that the noisy, unaligned data shown in Figure 4 (which naturally included many words that the child did not produce) may have obscured the fine-tuning of utterance complexity for those words that the child was able to produce.

Although this analysis shows the promise of the Speechome corpus (since analyses of CDS time-series data at the level of single words have never before been possible), it also

[3]Because words appear in multiple utterances (and more words appear, by definition, in longer utterances) absolute MLU in this calculation is biased upwards. Thus, we normalize the MLU curves by subtracting the MLU at the time of word birth.

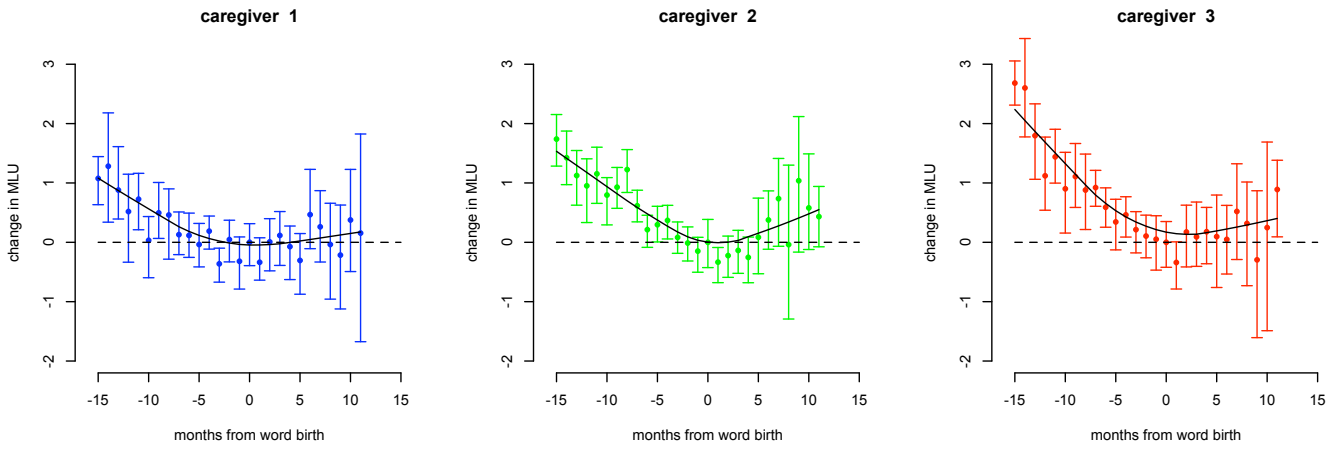

Figure 5: Mean caregiver MLU across words when the caregiver's MLU time series are aligned by word birth (zero on the X axis) and centered at the MLU at birth (zero on the Y axis). Each panel shows a different caregiver. Error bars are 95% confidence intervals computed by a non-parametric bootstrap and a lowess smoothing line is plotted in red for each curve.

reveals the limitations of even the current dataset. Without denser data it is impossible to look at coordination in the short window of time immediately before and after the child's production of a new word, and these short temporal dynamics may also reveal effects of fine tuning.

Conclusions

The Human Speechome Project represents a novel opportunity to explore hypotheses about the relationship between caregivers' speech and the linguistic abilities of the child. We found evidence that word frequency in CDS influences the child's age of acquisition for those words. We also found strong evidence for caregivers' modification of the length and lexical diversity of their utterances contingent on the child's linguistic ability. In addition, we found some support for a finer level of lexical tuning, the modification of utterance length on a word-by-word basis according to whether the child knows that word or not.

More generally, the current analyses constitute only a first look at an extremely rich dataset. As transcription progresses, we can look forward to enriching a number of the current analyses with more accurate assessments of the child's productive vocabulary and the short-term dynamics of caregivers' speech surrounding the first use of a word. Furthermore, the visual information contained in our database offers an unparalleled opportunity to explore more detailed questions about the interaction between linguistic and physical context in acquisition. Our hope is that through the power of this resource we may be able to make new progress on long-standing questions in child language development.

References

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

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D. J., Pethick, S., et al. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*, 1-185.

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

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

Hart, B., & Risley, T.(1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Brookes Publishing Company.

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

Lieven, E., Salomo, D., & Tomasello, M. (in press). Two-year-old children's production of multiword utterances: A usage-based analysis.

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

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.

Piaget, J. (1952). *The origins of intelligence in children*. International Universities Press.

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

Roy, B. (2007). *Human-machine collaboration for rapid speech transcription*. Unpublished master's thesis, Massachusetts Institute of Technology.

Roy, B., & Roy, D.(under review). Fast transcription of unstructured audio recordings.

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* (p. 2059-2064). Mahwah, NJ: Lawrence Earlbaum.

Snow, C., & Ferguson, C. (1977). *Talking to children.* Cambridge, UK: Cambridge University Press.

Tomasello, M.(1992). *First verbs: A case study of early grammatical development*. Cambridge, UK: Cambridge University Press.