Large-scale investigations of variability in children's first words

Rose M. Schneider
rschneid@stanford.edu
Department of Psychology
Stanford University

Daniel Yurovsky
yurovsky@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

First words are both an intimate moment between child and caregiver and an important step towards language. Across children, the timing of the first word varies widely, as does its semantic category and phonological complexity. Using a number of large new datasets of parent reports, we investigate patterns in children's first recognizable production of a word. We find that, contra conventional wisdom, more than 75 percent of children in our data produce a first word by their first birthday. In addition, children consistently produce more first words in certain semantic categories, across a wide range of ages. Finally, we create a predictive model of first word production using both input frequency and phonological complexity and find that both are important contributors to first word production. Throughout our analyses, we find that parental reports of a child's first word yield rich and consistent data on what is typically an unobservable dyadic moment.

Keywords: language acquisition

Introduction

A child's first word production is not only a highly anticipated and memorable event, but also a brief glimpse into the mind of an infant acquiring language, one of the most astonishing development processes. Over the course of their first years, children rapidly go from speechless infants to toddlers producing and learning language at an astounding rate (Fenson et al., 1994; Bloom, 2002). Marking the beginning of productive language, the first word is an important and measurable insight into a child's conceptual and linguistic development. Yet, in contrast to later milestones, children's first words are an intimate moment between child and caregiver, and are difficult for external observers to record or measure. Here we leverage large-scale data from parental reports to ask what children's first words reveal about two key issues in early language learning; the time-course of the emergence of language, and the relation of linguistic and conceptual development.

First, from a very early point in development, infants exhibit an aptitude for language, even showing a preference for infant-directed over adult-directed speech at 1 month (Cooper & Aslin, 1990). Over the course of the first year, infants are learning to recognize the distinctive sounds and word forms of their native language (Kuhl, 2004) and to segment these phonetic forms and words (Werker & Curtin, 2005). Additionally, by 6 – 9 months, many infants already show a tendency to look to matching pictures when they hear common nouns, suggesting early beginnings for form-meaning mapping as well (Tincoff & Jusczyk, 1999, 2012; Bergelson & Swingley, 2012). Infants' language comprehension abilities thus appear to be reasonably well-developed prior to 12 months. As early as 12 weeks, children begin producing the sounds of their native language in babble (Kuhl & Meltzoff, 1996), suggesting an early beginning to linguistic production. Conventional wisdom holds that word production typically begins around the first birthday, however. Is this early lag between comprehension and production real, or only apparent?

Second, what is the relationship between the words children learn and their conceptual development? Typically-developing monolingual children show correlations between some cognitive achievements and their language production; for example, acquisition of words about disappearance is correlated with comprehension of object permanence (Gopnik & Meltzoff, 1986). But at a larger scale, conceptual development appears to play a more limited role: 2 – 5-year-old international adoptees learning English for the first time show the same gross pattern of language acquisition as monolingual infants in terms of vocabulary composition (Snedeker, Geren, & Shafto, 2007). Additionally, there are also striking convergences in early words across very different cultural contexts (Tardif, 2007). Do patterns of first word productions—and their distribution across semantic categories—suggest any relationship between language acquisition and cognitive development?

Because very early language is difficult to observe in the lab, we leverage parent reports to learn about children's first words. A child's first word is highly memorable for parents, and many parents record this milestone in baby books. While recognizing that parents may not share all aspects of this standard, we define a true "first word" as the consistent use of a form to communicate a particular meaning, whether or not that form matches the adult target. Intuitively, we believe this is what parents tend to think they are reporting when they report first words, though they may occasionally be biased to construct consistency and communicativeness out of babble and coincidence.

As a scientific measure, parent report has both substantial disadvantages and real advantages. One issue with any self-report measure is that there is no way to validate participants' responses. Another complication is that parents may be biased observers, and interpret word-like babble as productive communication. Nevertheless, parent report is widely used as a measure of early child language, e.g. in the MacArthur-Bates Communicative Development Inventory (CDI), a vocabulary checklist that is both a reliable and valid measure of early vocabulary (Fenson et al., 1994, 2007) (although the re-

liability of the earliest ages of the CDI has been questioned; Feldman et al., 2000). On the other hand, self-reports are very easy to collect, making them ideal for large-scale investigations like the present study.

To address issues of bias in self-report, we gathered data from a number of sources. The first dataset source was the member families from of a local children's museum; these parents were an ethnically diverse population with a higher education level than the general population and a demonstrated interest in their child's development, likely leading to more engagement in their children's early language. Our second dataset was the Amazon Mechanical Turk parent population; this community is more diverse in terms of age, gender, education level, and socio-economic status (SES). Our third dataset came from parents in the psycholinguistic research community. We selected this population for its familiarity with the subject and because this community was most likely to have written records about first words. The data we received from all three of these surveys was generally very consistent, both within and across datasets, leading us to believe at very least that any bias in one was likely in operation across all three.

Our final dataset was drawn from Wordbank, an open repository of a large amount of CDI form data that aggregates across several samples including the updated CDI norming sample (Fenson et al., 2007). This dataset was chosen because the CDI data come from forms filled out with respect to a child's current productive vocabulary; thus this dataset lacks the retrospective reporting biases of our other surveys. Because the CDI contains a fixed set of words, it constrains the space of possible first words but also facilitates comparative analyses by reducing the space to a small, representative set.

Drawing on these datsets, we investigate the time-course of the emergence of productive language and potential factors that might lead to individual differences in linguistic development. First, we analyzed variability in the age of first word onset, finding that 75% of children were reported as producing a word prior to 12 months. Second, we asked whether the range of of first words varies with children's chronological age, allowing us to ask about the relationship between linguistic and conceptual development. This analysis yielded no measurable differences, indicating that linguistic factors—rather than conceptual ones—likely constrain the set of first words. Finally, we show that two specific linguistic factors, input frequency and phonetic complexity, both predict the words that children are likely to say first.

General Methods

Data for the study come from four datasets. Three of the four were surveys specifically designed for this study.

Dataset 1: Museum Member Survey

Participants We sent out a very brief survey on children's first words to subscribed members of a large local children's museum. We received responses for 502 children (215 female, 285 male, and 2 with no reported sex; M age = 11 mo., median = 10 mo.). Several responses were translated into English where possible; one response could not be translated and was excluded from further analysis.

Method Parents completed a web-based survey (created with JavaScript and HTML). The survey asked parents to report their child's first word (excluding "mama" and "dada"), the word referent, a description of the situation surrounding the first word, the child's age at time of utterance (10 mo. or younger, 11 mo., 12 mo., 13 mo., 14 mo.), the child's current age, and sex. Parents answered for only one child in this survey. We standardized responses and corrected obvious spelling errors; when the meaning of the word was not immediately apparent, we relied on the parent's description of the circumstances surrounding the word and/or the parent's classification of the word type.

Dataset 2: Amazon Mechanical Turk

Participants We recruited 1000 parents from Amazon Mechanical Turk to complete a more in-depth survey on their children's first words. We restricted the survey to parents in the United States. This survey allowed parents to answer for multiple children. We received responses for 1671 children (813 female, 858 male; M age = 10 mo., median = 10 mo.). Responses from 21 children were excluded from subsequent analyses because they had not yet spoken (M age = 2.7 mo., median = 2 mo.). Responses in other languages were translated into English where possible; one response was excluded. Caregiver education levels were highly diverse (Elementary = 3; Some high school = 15; High school = 166; Some college = 309; College = 346; Some graduate school = 26; Graduate school = 131; Total = 996).

Method This survey was an extended version of the Museum survey, allowing for input for multiple children, and asking the respondent to list their highest level of education, child's birth order, sex, first word (excluding "mama" and "dada"), word type, addressee of the first word, age at time of the word (0 – 24+ months), current age (0 – 18+ years), and both home and word language.

Data were handled as in Dataset 1. Due to the larger sample size, more phonological and morphological variations appeared. A final standardized form was selected, and the various original first word forms were recoded as that standardized form. For example, "Dog dog," "Doggy," "Doggie," and "Dogie" were all coded as "Dog." We occasionally relied on the parent's description of the situation in this coding process.

Dataset 3: Psycholinguists

Participants We sent out a brief survey on children's first words to subscribed members of a Psycholinguistics listserv. We received 52 responses from this survey (26 female, 26 male; M age = 11.16 mo., median 11 mo.).

Method Questions included on the survey were: The approximate phonological form of the first word, the age of utterance, when the parent recorded the word (if at all), the child's sex, the target word, the child's birth order (first or later born), and the child's current age. Data were handled similarly to Datasets 1 and 2.

Dataset 4: Wordbank

Participants At the time of our analysis, the Wordbank database contained 949 unique CDI Words and Gestures administrations. From these, we selected the 76 children whose

CDM	MTurk	Psycholinguists	Wordbank
Ball	**Dog**	**Up**	Baa Baa
Hi	**No**	**More**	**Uh–Oh**
Dog	**Ball**	**Hi**	Yum Yum
Uh–Oh	**Bottle**	**Cat**	Woof Woof
Duck	**Hi**	**Bye**	**Hi**
Car	**Bye**	-	Vroom
No	Kitty	-	This
Cat	Baba	-	Meow
Bye	**Cat**	-	**Bottle**
Up, **More**	Milk	-	**Ball**

Table 1: Top ten first words (excluding "mama" and "dada") from each of the four datasets we examined. Words repeated across datasets are bolded. Included are only words with more than 1 instance.

parents reported that they produced exactly one word (31 female, 45 male, M age = 10.63 mo., median = 11 mo.). Caregiver education levels were fairly diverse (Some high school = 4; High school = 24; Some college = 21; College = 17; Some graduate school = 1; Graduate school = 9).

Data preparation As responses were taken directly from the CDI, no data preparation was necessary.

Exclusion of "mama" and "dada"

While many parents reported that their child's first word was "mama" or "dada" (or some equivalent or variant), we excluded these children from our analyses. Parents may be motivated to hear these words very early in babble, even when the word is not being used in a meaningful or consistent way. Therefore, we stressed in our surveys that parents were to report their children's first word *other* than "mama" or "dada" to avoid this possibility and to detect a larger range of conceptual types. Additionaly in the MTurk dataset, we included a question asking whether the child's first word was "mama", "dada," or another first word. In total, 1112/1650 (67%) of children were reported to produce "mama" (N = 618) or "dada" (N = 489) first rather than another word (N = 543).

Analyses

Table 1 shows the top ten words from each dataset. Overall, there is substantial consistency across the four datasets, with "Hi" appearing in all four, and "Bye", "Ball", and "Cat" appearing in three. We use these four datasets to conduct three primary analyses. Analysis 1 examines the age of first production, Analysis 2 describes the semantic categories of these words, and Analysis 3 attempts to predict which words tend to be produced on the basis of phonological complexity and input frequency.

Analysis 1: Age of First Word

Despite evidence for very early word comprehension (Tincoff & Jusczyk, 1999, 2012; Bergelson & Swingley, 2012), conventional wisdom generally holds that first word emerges at around 12 months. However, a child's first word is almost exclusively heard by a parent or other caretaker. Is this reported lag between comprehension and production real or apparent?

Using data from a total of 2,279 children we plotted the cumulative probability of a child having produced a first word as a function of their age and dataset (Figure 1). Prior to 12 months, approximately 75% of children had produced a first word, across all four datasets. This result was strikingly consistent across datasets, despite significant variance in the tails.

Data from the Museum survey were truncated due to a "ten months or earlier" response option and showed the least age variability, with respondents modally choosing the earliest option. Data from Wordbank were also truncated due to the 8 month cutoff for the use of the CDI (as well as data sparsity in the latest ages); nevertheless, Wordbank data showed the earliest word productions. One possibility is this may reflect a bias towards reporting at least one word, given the process of going through the entire CDI checklist; another possibility is that seeing the checklist allows parents to more thoroughly consider their child's early productions.

Data from the MTurk survey showed a broader distribution of ages, perhaps due to the greater diversity (as well as larger size) of this sample. Some children were reported to be producing words implausibly early (e.g., 4 months). These responses are very likely (though not, to be fair, with absolute certainty) the result of reporting errors or biases. To estimate retrospective reporting biases, we regressed the mean age of first words against the time since the event in our largest dataset (MTurk), but did not find a significant relation, suggesting no biases of this type that we could measure. On the other end of the spectrum, some respondents reported first words appearing after 18 months, a timeline which might raise clinical concerns. Indeed, in a population as large as this one, there are almost certainly included some children with speech-language delays or other developmental disorders. Thus, this dataset is potentially valuable for estimating the right tail of the distribution in a diverse population.

Finally, the Psycholinguist dataset shows a relatively later and steeper onset of word production than the other three (though it still reaches 75% around 11 months). Given the high level of education of the respondents, it is likely that these children would have large early vocabularies (e.g. Hart & Risley, 1995, et seq.). On the other hand, the majority of these respondents recorded their child's first word at the time of production, decreasing concerns about retrospective report. Additionally, these respondents had training in psycholinguistics and were more likely to apply a more stringent standard (we shared our definition of a first word with respondents in the survey instructions). Thus we view the lack of very early respondents as prima facie evidence that first words before 9 months are rarer than our other surveys might lead us to believe.

In sum, we see some evidence for over-optimism (estimating first words earlier than we might expect) in a number of our datasets. Nevertheless, there was no evidence for retrospective reporting biases in our largest dataset; a more plausible account is that first words (whether optimistically or realistically detected) are a memorable event whose date and context are recalled well. In addition, despite differences in the tails, there was a striking convergence between datasets in suggesting that most children in our sample produced a first

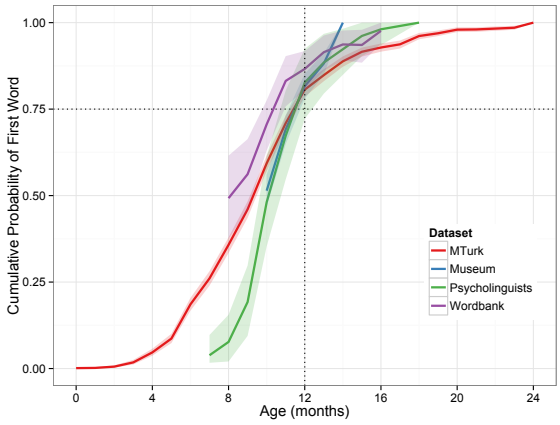

Figure 1: Cumulative probability of a child having produced her first word across development. In all datasets, more than 75% percent of children had produced their first word by their first birthday and more than half had produced their first word by 10 months. Shaded regions show 95% confidence intervals computed by non-parametric bootstrap.

word prior to their first birthday.

Analysis 2: Independence of Age and First Word

The variability in children's age of first production gives us a natural tool for asking about the relationship between conceptual and linguistic development. All things being equal between samples—which they are almost certainly not, since younger language producers might be more sophisticated conceptually as well—younger children should be less conceptually sophisticated and hence might produce words for a more restricted range of concepts. Alternatively, if the concepts that children most want to talk about are present early (Snedeker et al., 2007; Snedeker, Geren, & Shafto, 2012; Gleitman, 1990), we should predict no difference in the distribution of first words for older and younger children.

In our next analysis, we tested the hypothesis that the distribution of first words would differ between older and younger children. To avoid issues of data sparsity across word forms, we assigned words to the categories that appear on the CDI instrument (e.g., Animals, Games and Routines, Toys, People, etc.) and conducted our analysis over the category distribution of words (a loose proxy for their semantic distribution). We assigned CDI categories consistently across datasets for words that did not appear on the CDI word list. Ninety-one children were excluded because their first word could not be categorized.

Figure 2 shows the frequencies of the CDI categories split by age (at the first birthday) and grouped by dataset. Animals, Games and Routines, Toys, and People, followed closely by Food and Drink, were frequent first word categories. All of these seemed to be equally compelling as a first word for both early and later producers. Data for later speakers in Wordbank was sparse, because children were selected for this analysis when they were producing exactly one word, according to parental report on the CDI. Only 27 children produced only exactly word in this group.

Distributions of CDI category frequency across all 4 datasets were quite similar. ENTROPY ANALYSIS HERE? We constructed a linear model, and did not find an effect of dataset on CDI category distribution ($p > XX$). Although more children are included in the > 12 month group, the distributions of the CDI categories remain very similar.

In sum, despite producing a first word during different points in their conceptual development, both early and later producers in our sample chose to talk about the same semantic categories, and in many cases, the same things. This finding suggests that first words tend to reflect concepts that are available early (at least to those children who have the wherewithal to talk about them). Why then do children consistently pick certain words? In the next analysis, we examine the role of input frequency and phonological complexity in determining which words are predicted.

Analysis 3: Predicting First Words

Our previous analyses found a high degree of consistency among children's first productions. Independent of their age, children seem to produce the same first words. Why these words? Our analyses, along with those of Snedeker et al. (2012) suggest a degree of independence between conceptual and linguistic development. Thus, we hypothesize that children's first words should be constrained by their linguistic input and their speech production abilities. To say her first word, a child must minimally have been exposed to that word, and also be able to pronounce it. We thus ask whether children's first words are predicted by two factors known to predict the acquisition of words more broadly: input frequency, and phonetic complexity (Morgan & Demuth, 1996; Goodman, Dale, & Li, 2008).

Method Each of our datasets consists of a set of words that children produced, and the number of children who produced each of these words. Our goal in the final analysis is to determine both why some of these words were first words more frequently than others (e.g. "dog" vs. "asleep"), and also why some words were never first words at all (e.g. "animal"). Because the set of words that were never produced is infinite, we needed to constraint our set of candidate first words to a small, representative, finite set. For this reason, and to ensure fair comparison across datasets, we restricted our set of words to the 385 words that appear on the CDI Words and Gestures form. We thus selected the subset of first words in each of our datasets that appeared on the CDI, and asked how many of our children produced each.

To estimate the frequency with which children hear each of these words, we counted the number of times each appeared in CHILDES—a large corpus of parent-child interactions (MacWhinney, 2000). In order to ensure a representative sample, we counted the number of appearances of each word in a child's mother's speech across all of the corpora in the North American subset of CHILDES. These frequencies were then log-transformed and used to predict first words.

To estimate phonetic complexity, we chose a simple, theory-independent measure: number of phonemes. For each of this same subset of words, we queried the MRC Psycholinguistic Database (Wilson, 1988). Number of phonemes is an imperfect measure of phonetic complexity—it misses differ-

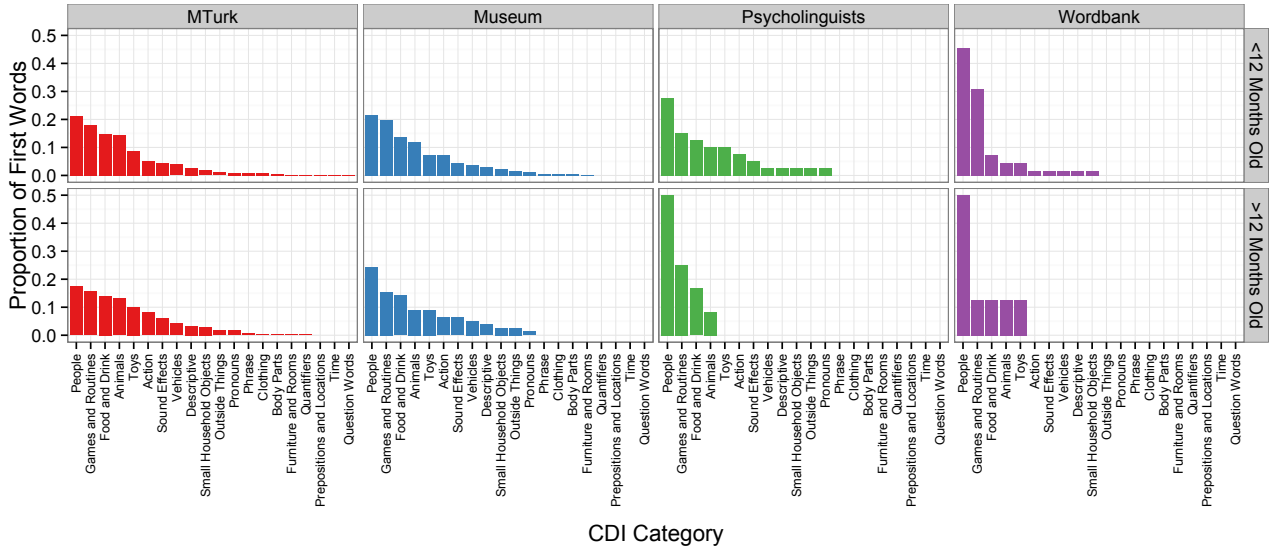

Figure 2: Proportion of children's first words falling into each CDI category. The datasets showed a high degree of consistency, with most first words referring to animals or games and routines. These distributions were highly consistent between older and younger children, suggesting that first words are driven by linguistic rather than conceptual factors.

ences in articulatory complexity that contribute to the relative difficulty of producing different words (e.g. "truck" vs. "bunny")—but it does capture much of the variability among the CDI words.

To predict the number of observations of each of a set of a categorical outcomes, the standard statistical model is Poisson regression. However, Poisson regression behaves poorly in many datasets because empirical count distributions violate its assumptions in two ways: Their variance is greater than expected (over dispersion), and too many of them are zero (zero inflation). To adjust for these violations, we used a hurdle model (Mullahy, 1986). This model predicts the number of observed counts through a combination of two processes: a binomial threshold (hurdle) that first determines with a count is zero or greater, and then a second component which determines the size of the count if it is non-zero. Because the datasets were of such different sizes, we fit a separate hurdle model to each and examined consistency in the estimated parameters across datasets.

Results and Discussion Across datasets, input frequency and phonetic complexity consistently predicted the number of children who produced each word as their first word. As we hypothesized, in almost all cases candidate words were more likely to be first words if they were higher frequency in children's input, and if they had fewer phonemes (Figure 3). In conjunction with the analyses above, these results suggest a high degree of consistency in children's first productions, independent of conceptual development, and dependent instead on linguistic input and speech production fluency.

General Discussion

What can children's first productive utterances reveal about their conceptual and linguistic development? Utilizing large-scale data from parental reports of first words, we conducted three major analyses on the time-course of productive language emergence, distributions of conceptual categories across early and later first words, and some factors that play a role in predicting first word production. Interestingly, we found that 75% of children produce a first word prior to 12 months, the conventionally cited age. Furthermore, we found that children's first words generally reflect the concepts that are available early in development, suggesting some independence between conceptual and linguistic development. Finally, we suggested that two factors – input frequency and phonetic complexity – consistently predicted the number of children producing a given word as their first utterance.

Despite the degree of consistency we observed in these datasets, it is important to again acknowledge the limitations and drawbacks of data drawn from parental self-report, and the same caveats that present themselves with any data collected by self-report must be observed here as well. We attempted to counter these issues by collecting data from a number of diverse samples, and by providing explicit definitions of what consituted a first word. Additionally, we did not find evidence of overt biases in reporting very young first word productions, and we observed similar patterns of productions across all four datasets.

Our findings suggest that the lag between comprehension and production in the first year of life may be shorter than traditionally thought. Additionally, conceptual and linguistic development seem to exhibit a certain degree of independence, with children early and later in development producing words about concepts that are available to them early. It is likely therefore, that linguistic rather than conceptual factors may constrain the set of possible first words and give rise to some words being consistently produced as first utterances.

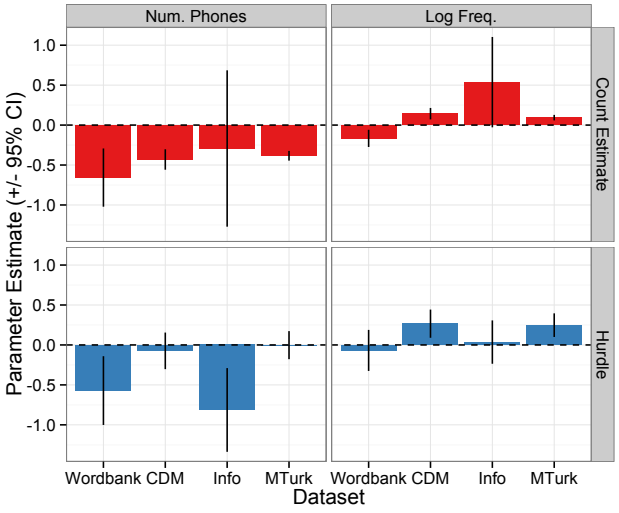

Figure 3: Parameter estimates for hurdle models predicting children's first words. Models showed a high degree of consistency across datasets: first words tend to be higher frequency and have fewer phonemes. Intercepts are omitted for clarity.

Acknowledgements

Thanks to Ally Kraus for assistance with survey design and Jenni Martin and Rick Berg at Children's Discovery Museum of San Jose for help with Dataset 1.

References

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the macarthur communicative development inventories at ages one and two years. *Child development*, *71*(2), 310–322.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *Macarthur-bates communicative development inventories*.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*(5).

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 3–55.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Gopnik, A., & Meltzoff, A. N. (1986, August). Relations between semantic and cognitive development in the one-word stage: The specificity hypothesis. *Child Development*, *57*(4), 1040-1053.

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

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831–843.

Kuhl, P., & Meltzoff, A. (1996, 2438). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, *100*(4), 2425.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. third edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Morgan, J. L., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*(3), 341–365.

Snedeker, J., Geren, J., & Shafto, C. (2007). Starting over: international adoption as a natural experiment in language development. *Psychological science*, *18*(1), 79.

Snedeker, J., Geren, J., & Shafto, C. L. (2012). Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children. *Cognitive Psychology*, *65*(1), 39–76.

Tardif, C. (2007). Cross-linguistic differences in children's early word acquisition. *Advances in Psychological Science*, *3*.

Tincoff, R., & Jusczyk, P. W. (1999). Some Beginnings of Word Comprehension in 6-Month-Olds. *Psychological Science*, *10*, 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

Werker, J. F., & Curtin, S. (2005). Primir: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10.