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

The first word, an intimate moment between child and caregiver, exhibits a tremendous amount of variability in semantic categorization, phonological complexity, and age of onset. Through several large datasets of parental report of children's first words, we investigate patterns in first word production. In three analyses, we explore the time course and distribution of children's first recognizable language productions. We find that, contra conventional wisdom, more than 75 percent of children in our datasets produce a first word by their first birthday. In our second analysis, we find that children consistently produce more first words in certain semantic categories. Finally, we take all the unique occurrences of words across the datasets, and try to predict first word production via parental input taken from the CHILDES corpus and the words' phonetic probabilities. Overall, we find that parental report of a child's first word yields rich and consistent data on what is typically an unobservable dyadic moment, and that consistencies in first word production across development may indicate a close relationship between conceptual and linguistic development.

Keywords: language acquisition

Introduction

The emergence of language in infancy is among the most astonishing developmental 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 verbal language, a child's first word is an important and measurable insight into what a child is willing and able to talk about at that point in their development. Yet, in contrast to later milestones, children's first words are an intimate moment between child and caregiver that is 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.

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 (Bergelson & Swingley, 2012). Infants' language comprehension abilities appear to be reasonably well-developed prior to 12 months. However, conventional wisdom holds that language production typically begins around the first birthday. Is this 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 exhibit some particular relations between certain cognitive developments and language productions; for example, acquisition of words about disappearance is correlated with the comprehension of object permanence (?, ?). In contrast, 2 and 5 year–old international adoptees newly learning English show the same pattern of language acquisition as monolingual infants, both in time course and pattern of productive utterances (Snedeker, Geren, & Shafto, 2007). This dissociation suggests that at older ages conceptual development may interact very little with word learning, but what about at younger ages with general referential language acquisition? Thus, when trying to untangle conceptual and linguistic development in infants, the first word is a very important as a measurable developmental milestone and as a more concrete indicator of a child's developmental state. Do patterns of first word productions 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. As a scientific measure, however, parent report has both drawbacks and 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. 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). Nevertheless, the reliability of the earliest ages of the CDI has been questioned (Feldman et al., 2000).

To address the issues of self–report, we gathered data from a number of sources. We specifically targeted different populations in our selections of data sources. The diversity of these four datasets encouraged an representative sample of children's first words, with each data source contributing its own set of advantages and drawbacks to the complete dataset.

The first dataset source, parents subscribed to The Children's Discovery Museum (CDM) mailing list, was chosen for collection because these parents were an ethnically diverse population with a higher education level than the general population, potentially leading to more accurate responses. For our next data source we targeted the Amazon Mechanical Turk (MTurk) population. This community is more representative of the general population in terms of age, gender, education level, and socio-economic status (SES).

To complement our data from the general population, we next collected first word data from parents in the Psycholinguistic community. This population was specificaly selected for its familiarity with the subject as well as for highly accurate diary records. However, this population ended up being very small (N = 58). The data we received from all of these surveys was generally very consistent, both within and across datasets, indicating that parents responded appropriately. In addition to yielding reliable data, our first word surveys were highly effective, and were completed quickly and with minimal effort by many parents.

Our final dataset, drawn from `Wordbank`— an open repository of CDI data. This sample was chosen because these parent-report measures come from forms filled out closer in time to when the child produced their early words. Because the CDI contains a fixed set of words, it constrains the space of possible first words to a finite set. This limits the range of possible first words we can observe, but also facilitates comparative analyses by reducing the space to a small, representative set of possible words.

Drawing on these datsets, we investigate the timecourse 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 percent of children produce 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. We found no difference, indicating that linguistic factors rather than conceptual ones constrain the set of first words. Finally, in we show that two specific linguistic factors—input frequency and phonetic complexity—*do* predict the words that children are likely to say first.

General Data Collection Methods

Data for this study is comprised of four different datasets, each obtained from a different source. Three of the four datasets were drawn from surveys specifically designed for this study. The last dataset contains data from `Wordbank`, an online repository of data from the MacArthur–Bates Communicative Development Inventories (CDI), a widely used parent-report vocabulary checklist (Fenson et al., 2007).

The first dataset source, parents subscribed to The Children's Discovery Museum (CDM) mailing list, was chosen for collection because these parents were an ethnically diverse population with a higher education level than the general population, potentially leading to more accurate responses. For our next data source we targeted the Amazon Mechanical Turk (MTurk) population. This community is more representative of the general population in terms of age, gender, education level, and socio-economic status (SES).

To complement our data from the general population, we next collected first word data from parents in the Psycholinguistic community (Info). This population was specifically selected for its familiarity with the subject as well as for highly accurate diary records. However, this population ended up being very small (N = 58).

Our final dataset, drawn from `Wordbank` was chosen because data points came from the MB-CDI, which is typically filled out closer in time to when the child produced her first word. A potential confound of the MB-CDI is a possible demand characteristic encouraging parents to report their children producing some word, as well as containing a somewhat limited set of potential words.

Overall, these methods of parental report yielded consistent and rich data with minimal time investment on behalf of the parent. However, as with most self–report measures, our surveys did have some disadvantages. While in every data collection we tried to stress that a first word was defined as a consistent set of sounds referring to the same referent across many contextual frames, we have no way of validating the parental report. Another issue was standardization of the child's first word, especially in the MTurk dataset, discussed below.

Dataset 1: Children's Discovery Museum Survey

Participants

We sent out a brief survey on children's first words to subscribed members of a large local children's museum. We received 502 responses to our survey (215 female, 285 male, and 2 with no reported sex; M age = 11 mo, median = 10 mo). Due to the diversity of the San Jose community, several of the first word responses were not in English. Responses were translated into English where possible. Responses that could not be translated were excluded from further analysis (N = 1). Guardian education level was not available for this population.

Method

Parents completed a brief web–based survey (created with JavaScript and HTML). The survey asked parents to list their child's first word (excluding "mama" and "dada"), what they thought word referred to, 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.

Parents' responses were standardized for ease of analysis. Data cleaning involved fixing obvious spelling errors. When the meaning of the word was not immediately apparent, the researcher 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 an updated 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 1671 responses (813 female, 858 male; M age = 10 mo, median = 10 mo). 21 children were excluded from subsequent analyses because they had not yet spoken (M age = 2.7

mo, median = 2 mo). Responses were translated into English when possible and required. Responses that were not able to be translated were excluded from further analysis (N = 1). Guardian education levels were highly diverse (Ns as follows: Elementary = 3; Some high school = 26; High school = 308; Some college = 525; College = 553; Some graduate school = 42; Graduate school = 26).

Method

This survey was an extended version of the previous one. The survey allowed for input for multiple children, and asked parents to list their highest education level, child's birth order, sex, first word (excluding "mama" and "dada"), word type, addressee of the first word, word age (0 - 24+ months), current age (0 - 18+ years), word language, and home language. Responses were validated as the survey was completed, reducing the likelihood of erroneous or false responses.

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 became that standardized form. For example, "Dog dog", "Doggy", "Doggie", and "Dogie" were all treated as "Dog" in the standardized form. We occasionally had to rely on the parent's description of the situation of the word occurrence to inform our decisions.

Dataset 3: Contemporary Psycholinguist Diary Studies

Participants

We sent out a brief survey on children's first words to subscribed members of a Psycholinguist 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 the utterance, when the parent recorded this (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 MCDI Words and Gestures administrations. From these, we selected the 76 children whose parents reported that they produced exactly one word (31 female, 45 male, M age = 10.63 mo, median = 11 mo). Guardian education levels were fairly diverse (Ns as follows: Some high school = 4; High school = 24; Some college = 21; College = 17; Some graduate school = 1; Graduate school = 9).

Data preparation

Because responses were taken directly from parental reports on the MB-CDI, no data preparation was necessary.

Analyses

The emergence of language in infancy has long been the focus of research and discussion. Noun comprehension has been shown to occur prior to production, indicating a conceptual understanding of label-referent mapping at 6-9 months (Tincoff & Jusczyk, 1999, 2012; Bergelson & Swingley, 2012). As early as 12 weeks, children begin producing the sounds of their native language /citekuhl1996, suggesting an early beginning to linguistic development. Yet the first word is not typically expected until around 12 months. Does a lag between comprehension and production exist, and what is the relationship between the words that children learn and their conceptual development?

In analyzing this data, we are interested first in understanding the timeline of the emergence of productive language and the factors predicting this emergence in individuals. We explore whether productive language lags comprehension with an analysis of children's ages at the time of their first words. We find that 75% of children produce a first word by 12 months as well as evidence of a wide age range in which the first word is typically uttered.

We next turn to the question of why some children are producing language earlier than others, and if this reveals anything about the relationship between conceptual and linguistic development. First, we compare distributions of semantic (CDI) categories in first word productions in children older and younger than 12 months. If early producers' CDI category distributions are significantly different than later producers', this could indicate a loose coupling of linguistic and conceptual development. However, we find that both before and after the first year, children's first words tend to come from the same conceptual categories.

We do not find that later producers are sampling words from a different or more diverse range of conceptual categories, but instead find that children across development are speaking about the same kinds of things, and often even the same things. What makes these words more likely to be produced as a first word, if they do not significantly differ conceptually? In our final analysis, we examined both the phonetic probability of individual words within CDI categories, as well as their correlation with parental input in an attempt to determine why some words are consistently produced before others. The results of this analysis indicate that fewer phonemes and increased input frequency within CDI categories predict first words production.

Overall, we find that if a lag between comprehension and general referential production exists in development, it is much shorter than previously thought, and most likely driven by developing linguistic capacities and not by a cognitive shift. Furthermore, our results suggest that conceptual and linguistic development are closely linked, and although the age of onset for productive language may vary across children, the semantic referents of first words are likely to be fairly consistent. Finally, we suggest that first words within semantic categories are predicted by phonetic complexity and input frequency.

Exclusion of "mama" and "dada"

While many parents reported that their child's first word was "mama" or "dada" (or some variant), we excluded these chil-

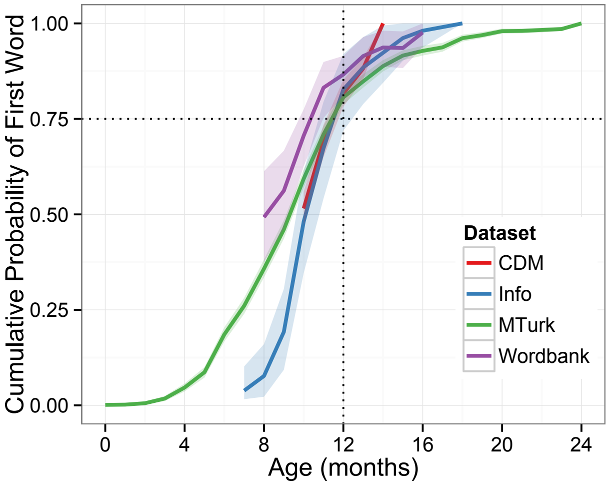

Figure 1: Cumulative probability of a child having produced her first word across development. In all datasets, more than 75% percent of children had produced their first word by their first birthday and more than half had produced their first word by 10 months. Shaded regions show 95% confidence intervals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

dren 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 the possibility of skewing the data. After pilot testing on Amazon Mechanical Turk, we added another question asking whether the child's first utterance was "mama", "dada", or another first word. 1112 parents reported their children producing "mama" (N = 618) or "dada" (N = 494) rather than another word (N = 559) as a first utterance.

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 and can be very difficult for an external observer to measure. Is this reported lag between comprehension and production real or apparent? Given the possibility of very early noun comprehension (Tincoff & Jusczyk, 1999, 2012; Bergelson & Swingley, 2012), we wished to explore the development of first word production, especially prior to 12 months.

Across the 4 datasets, we grouped data by age and by dataset, $N_{total} = 3173$. Twenty–one children were excluded for not having spoken yet ($M_{age} = 2.7$ mo., median = 2 mo.). We then plotted the cumulative probability of a child having produced a first word as a function of their age (Figure 1).

Prior to 12 months, approximately 75% of children have produced a first word, and we see a gradual but consistent increase from 12–24 months. The plot of the Mechanical Turk data, the largest dataset, is most likely the most representative of child word production, while the other datasets asymptote fairly quickly between about 14 and 17 months. These data suggest that production may not lag comprehension as much as previously thought. To discern whether there might be a bias on behalf of parents of older children to report a younger first word, we ran a (TEST), but did not receive a significant result ($p > XX$).

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.

(Snedeker et al., 2007; Snedeker, Geren, & Shafto, 2012; Gleitman, 1990)

Why are some children producing language earlier than other children, and what does this reveal about the relationship between their conceptual and linguistic development? Previous work has established that the CDI categories are not equally represented in first word production (Fenson et al., 1994). When children produce early language ($<$12 months), are they choosing to produce conceptually different words than children who speak later? Or, are conceptual and linguistic development more closely linked, with the same semantic categories appearing as first words regardless of age of utterance? In our next analysis, we examined the CDI category frequencies of children's first words to explore whether very early first words differ conceptually from later–produced words.

In this analysis, CDI categories were assigned by the researcher based on the CDI parental report form, and validated across datasets. In instances where words did not appear on the CDI, categories were assigned ad hoc and validated across datasets for consistency. Ninety–one children were excluded because their first word was unable to categorized. We grouped the data by dataset, and performed an age split at 12 months. Figure 2 shows the frequencies of the CDI categories split by age and grouped by dataset.

Distributions of CDI category frequency across all 4 datasets were quite similar. 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 category representations remain very similar.

Animals, Games and Routines, Toys, and People, followed closely by Food and Drink are frequent first word categories, and seem to be equally compelling as a first word for both early and later producers. Data for later speakers in Wordbank is 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.

Despite producing a first word during different points in their conceptual development, both early and later producers are choosing to speak about the same semantic categories, and in many cases, the same things (Figure 2). This would suggest that children's linguistic and conceptual development are more intimately related, and the production of a first word

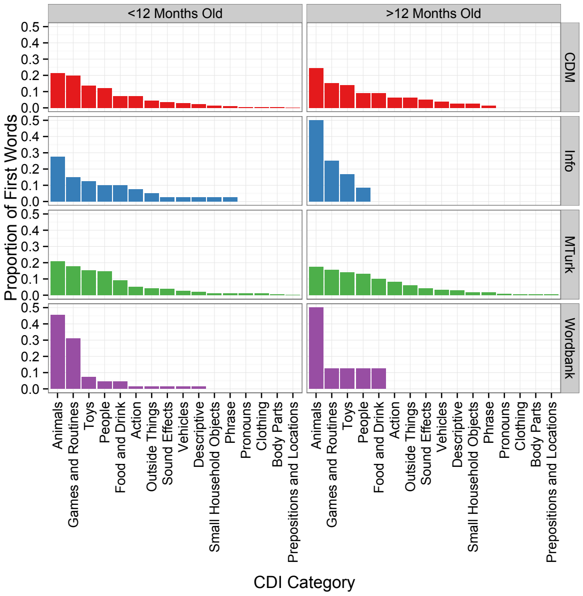

Figure 2: Proportion of children's first words falling into each CDI category. The datasets showed a high degree of consistency, with most first words referring to animals or games and routines. These distributions were highly consistent between older and younger children, suggesting that first words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

CDM	MTurk	Info	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

Table 1: Top 5 first words (excluding "mama" and "dada") from each of the four datasets we examined.

is possibly not the result of a cognitive shift, but an ability to produce and form words.

However, out of all the sets of possible first words, why are children consistently choosing to produce words predominantly from these semantic categories? What is tipping the balance in favor of these particular words? Some potential factors are frequency of parental input, the phonetic probability and ease of production. In the next analysis, we examine all of these within CDI categories to attempt to untangle what causes the measure of consistency we observe in first word productions.

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 For each of our datasets, our data consist 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 in 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. We thus chose the subset of 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 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—e.g. it misses differences 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

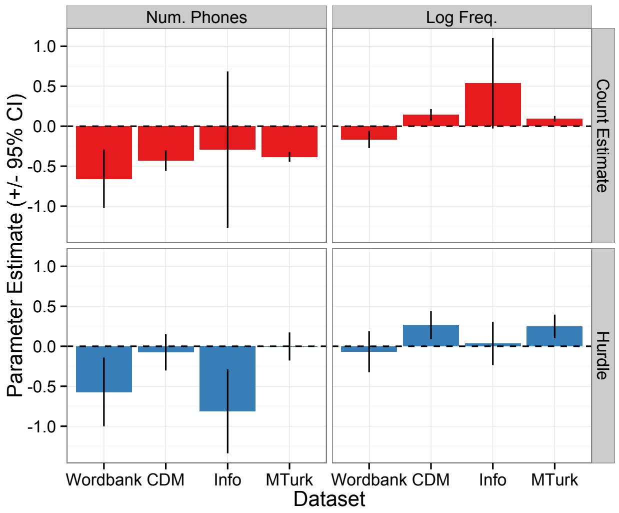

Figure 3: Parameter estimates for hurdle models predicting children's first words. Models showed a high degree of consistency across datasets: first words tend to be higher frequency and have fewer phonemes. Intercepts are omitted for clarity.

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

Acknowledgements

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., Choi, S., & Baumberger, T. (1996). Cross-linguistic differences in early semantic and cognitive development. *Cognitive Development*, *11*(2), 197–225.

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

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.

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.