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 caretaker, 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, including the age of onset, distribution of MB-CDI categories, and first words in relation to parental input and phonological complexity. In three analyses, we explore the timecourse 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

What can children's first words reveal about the process by which humans acquire language? Over the course of the first years, children rapidly go from speechless infants to toddlers producing and learning language at an astounding rate (citation about vocabulary acquisition here). The first word marks the beginning of productive verbal language, and is an important and measurable insight into what a child is willing and able to talk about at that point in their development.

Comprehension of language precedes its production for infants. Research from Bergelson and Swingley (2012) has suggested that as early as 6 months many children already understand what many common nouns mean, instead of just recognizing and segmenting speech sounds (cite bergelson and swingley here). Cognitively, children are capable of successfully mapping a label to an object very early; however, productive language emerges later, and goes through intense phonological and morphological changes throughout development (citation here). If children are able to grasp the meanings of words from a very early age, why does a lag exist between comprehension and production?

The first word is the start of productive linguistic capabilities, and is an important method for studying language acquisition. While there has been much research on the development of language over the course of the first few years, there has been less research focusing specifically first words. First word occurences are discrete and instantaneous events which almost exclusively involve only the child and parent, and are almost impossibly to observe in a lab setting. However, despite the scarcity of data, first word productions are important because they mark the very beginnings of the intersection of cognitive and linguistic capabilities.

Although a child's first word is unlikely to occur in an experimental setting, almost every child produces a first word, and this event is highly memorable for parents. Therefore, we explored first words on a large-scale through parental self report across 4 separate populations. This measure of children's first words has obvious drawbacks and advantages. One issue with any self-report measure is that there is no way to validate participants' responses. Another complication is whether parents were reporting the first true occurrence of a word, or merely babble. However, the data we received from these surveys was generally very consistent, both within and across datasets, indicating that parents responded appropriately. In addition to yielding consistent data, our first word surveys were highly effective, and were completed quickly and with minimal effort by many parents. To counter the potential issues presented by using a self-report measure, we were very careful in selecting our participant populations.

Using data collected from 4 different populations, we investigate the timeline of the emergence of productive language and some of the factors that may predict individual differences. Specifically, we look at the timecourse of first word production, and find evidence for word production prior to 12 months. An analysis of differences in conceptually different referents of first words for early versus later producers shows that older speakers do not first speak about semantically different things. Finally, we turn to factors that might affect first word production such as frequency of parental input and phonetic probability, and find ... something.

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 (MB-CDI), a widely-used parent-report vocabulary checklist (Fenson et al., 2007).

We specifically targeted different populations in our selections of data sources. The diversity of these four datasets encouraged an accurate representation of a child'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 he Amazon Mechanical Turk(MTurk) population. This community is diverse in terms of age, gender, education level, and socio-economic status (SES), and was chosen because it is more representative of the general population.

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 their highly accurate diary records. However, this population ended up being very small (N = 58).

Our final dataset, drawn from Wordbank, was chosen because datapoints came from the MB-CDI, which is typically filled out closer in time to when the child produced her first word, potentially yielding more accurate reports. A potential confound of the MB-CDI is that it may have a 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 gender; 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 were not able to be translated were excluded from further analysis (N = 1). Guardian education level was not available for this population.

Methods

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 their gender. Parents answered for only one child in this survey.

Data preparation

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).

Methods

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 preparation

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).

Methods

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 processing

Data were handled similarly to Datasets 1 and 2.

Dataset 4: MB-CDI Wordbank

Participants

From all available data on Wordbank, we selected 76 children whose parents reported as producing 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).

Methods

Data preparation

Analyses

The process by which children come to produce language has long been the focus of research and discussion. Noun comprehension has been shown to occur prior to production (?, ?, ?, ?), indicating that infants can successfully map words to objects. Children also begin producing the sounds of their native language fairly early in the form of babbles (citation here). However, the first word does not emerge until later (citation here). What are conceptual and linguistic developmental changes that occur over this time, and how are they related both to each other and to the appearance of the first word?

In analyzing this data, we are interesting in understanding the timeline of the emergence of productive language and the factors predicting this emergence in individuals. We first look at the age at which children's first word emerges to assess variability in onset of productive language. We find that 75% of children produce a first word by 12 months as well as evidence of a wider age range in which the first word is typically uttered.

Why are some children producing language earlier than others, and are these words significantly different semantically from later-produced words? Such a difference could indicate a loose coupling of linguistic and conceptual development, whereas if preference for certain semantic categories suggests that children at different stages of development are sampling first words from the same conceptual categories. We next wished to compare MB-CDI category distributions in early and later producers ($<$12 months and $>$ 12 months) to explore whether early producers' words are drawn from conceptually different categories than later producers'.

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 next analysis, we examined both the phonetic probability of individual words within CDI categories, as well as their correlation with parental input.

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 children from our analyses. Parents are obviously 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" or "dada", and found that 1112 parents reported their children producing "mama" (N = 618) or "dada" (N = 494) other than another first word (N = 559).

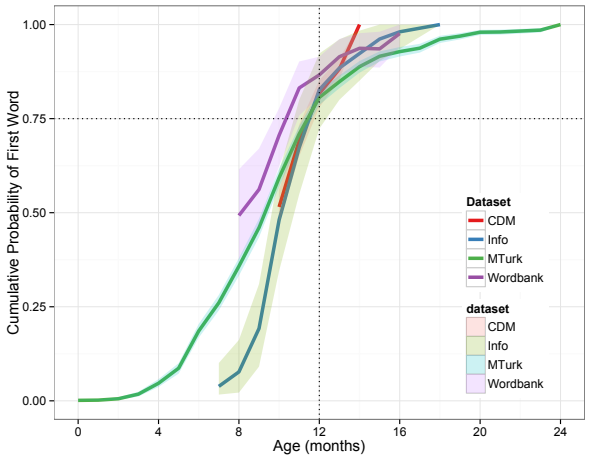

Figure 1: Age graph showing the cumulative probability of producing a word as a function of age.

Age

Conventional wisdom generally cites 12 months as the harbinger of the first word (Citation here). However, a child's first word is almost exclusively heard by a parent or other caretaker, and next to impossible to capture in a lab setting. However, that milestone is subject to the extreme variability in early language production (?, ?) and the potential pitfalls of parental memory (as discussed above). Given the possibility of very early noun comprehension (?, ?, ?, ?), 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).

Before 12 months, approximately 75% of children have produced a first word, and we see a gradual but consistent increase as the child ages as expected. The plot of the Mechanical Turk data, the largest dataset, is most likely the most representative of child word production, while the other datasets asymptope fairly quickly between about 14 and 17 months. These data suggest that children are speaking before 12 months relatively frequently. 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

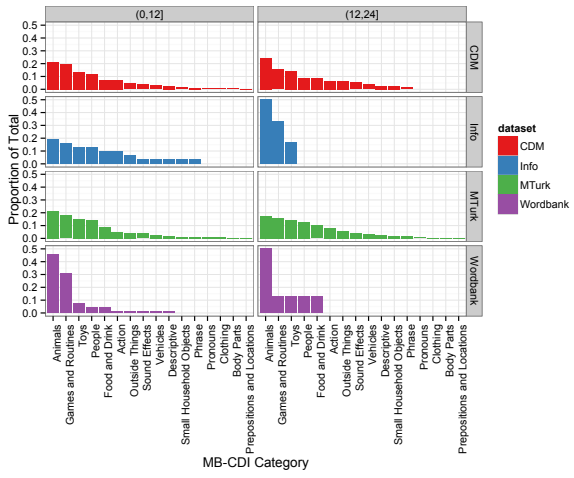

Figure 2: Histogram of CDI category frequencies across datasets.

significant result ($p > XX$).

Why are some children producing language earlier than other children, and what does this reveal about the relationship between their conceptual and linguistic development? When children produce early language (<12 months) as opposed to a later one, are they choosing to produce conceptually different words than children who speak later? Or, are conceptual and linguistic development more closely linked? 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.

CDI Categories

Previous work has established that the CDI normative categories are not equally represented in first word production (?, ?). If early word producers have a different distribution of CDI categories, this might suggest that linguistic development is loosely related to conceptual development, and that a later producer is sampling from a larger array of words that are conceptually available to her.

In this analysis, CDI categories were assigned by the researcher based on the MB-CDI parental report form, and validated across datasets. In instances where words did not appear on the MB-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 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 almost identical.

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 MB-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. This would suggest that children's linguistic and conceptual development are more intimately related, and the production of a first word is possibly not just the result of a cognitive shift, but an ability to produce and form words.

CDM	MTurk	Info	Wordbank
Ball (n = 51)	Dog(n = 110)	Up (n=3)	Baa Baa (n=16)
Hi (n = 29)	No (n = 108)	More(n = 3)	Uh-Oh (n = 7)
Dog (n = 29)	Ball(n = 102)	Hi (n = 2)	Yum Yum(n = 3)
Uh-Oh(n = 21)	Bottle(n = 76)	Cat (n = 2)	Woof Woof(n=3)
Duck (n = 16)	Hi (n = 57)	Bye (n = 2)	Hi (n = 3)

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, and potentially just relevance to a developing child. 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 variability.

- Describe analyses - Details of stats- - Differences across studies Consistent with previous work, we observe first words being drawn from certain CDI categories over others. We find that Animal, Games and Routines, Toy, and People words are more frequently first productions for both older and younger children, and that these patterns seem to be consistent across development.

These data support previous findings on CDI category analyses of early language production - Data suggests

Input Frequencies

- Question - Describe analyses - Details of stats - Differences across studies - Data suggests

Discussion

Acknowledgements