Large scale investigations of variability in early language production

**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. 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% of children in our datasets produce a first word by their first birthday. In our second analysis, we find that older children produce more 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. 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 the words children produce first may be affected by semantic differences as well as by parental input.

**Introduction**

Definitely need input on this, but also need to have another discussion about the claims we want to make based on having 4 (complete) datasets.

Intro: emergence of language is cool.

Paragraph: First words in comprehension – Bergelson & Swingley

Paragraph: Why are we interested in first productions?

Paragraph: parent report – advantages and disadvantages

Paragraph: summary of our wrok

**General Data Collection Methods**

Data for this study is comprised of 4 different datasets, each obtained from a different source. Three of the 4 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, a widely-used parent-report vocabulary checklist (Fenson et al., 2007).

**Survey 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.

Parents completed aThe mdy thought≤momomomo≥mo

**Data processing**

As the input for the child’s first word was free response, parents’ responses were standardized for ease of analysis. Data cleaning involved fixing obvious spelling errors; however, some responses were transliterated and representative of the phonetic pronunciation of the child’s first word. In instances when the intended word was obvious, the standardized first word became the final and correct pronunciation. 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, and the standardized word was listed as the final and correct form. Words for which this procedure was not possible were excluded from further analyses (N=XYZ).

**Survey 2: Amazon Mechanical Turk**

**Methods**

Our survey in study 2 was an extended version of our previous survey. The survey was programmed in JavaScript and HTML, and posted to Amazon Mechanical Turk externally. This survey allowed for input for multiple children, resulting in our large data set. The Mechanical Turk survey asked parents to input the highest education level of the mother (or primary caretaker), the birth order of the child they wished to answer for, the child’s gender, the first word (excluding “Mama” and “Dada”), the word type, the addressee of the first word, the word age (0 – 24+ months), the current age (0 – 18+ years), the word language, and the home language. Responses were validated as the survey was completed, reducing the likelihood of erroneous or false responses.

**Participants**

We recruited 1000 parents from Amazon Mechanical Turk to complete an updated survey on their children’s first words. To minimize the potential for errors in translation of first words, we limited the availability of the survey to Turk parents in the United States. As the survey in study 2 allowed the parent to answer the survey for multiple children, we obtained 1671 data points. Approximately 21 children were excluded from subsequent analyses because they had not yet spoken. Of the 1671 children listed in the responses, 813 were female, and 858 were male (mean age = 10 months, median age = 10 months). Although the HIT was restricted to the United States, several of the first words were in another language. These responses were not excluded, but the standardized forms of the word were translated into English when possible.

**Data Processing**

Data processing for survey 2 was very similar to survey 1. However, due to our additional questions about language, English translations for the standardized words were much more readily available. Because of the large sample size, many more phonological and morphological variations of first words were given. In these cases, a final standardized form was selected, and the various original first word forms became that standardized form. For example, parents listed “Dog dog”, “Doggy”, “Doggie”, and “Dogie” as the child’s first word; these were all treated as “Dog” in the standardized form. However, as discussed below, there was still some ambiguity in standardizing word forms, and we occasionally had to rely on the parent’s description of the situation of the word occurrence to inform our decisions. The most notable example in this dataset was “Baba” as the original first word response. “Baba” as an original first word had 80 occurrences, but according to the parent’s descriptions very often referred to “Ball,” “Bottle,” or “Grandma.” When the intent of the utterance was clear in the parent’s response, the appropriate standardized form of the word was listed. However, when this was not possible, the standardized form remained “Baba”, resulting in 37 “Baba” standardized first words.

**Survey 3: Contemporary Psycholinguist Diary Studies**

**Methods**

We distributed a short survey via email to a Psycholinguist mailing list. Participants were able to complete this survey more than once for multiple children. Questions included on the survey were: The approximate phonological form of your child’s 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), the child’s current age.

**Participants**

We received 52 responses from this survey (26 female, 26 male, mean age = 11.16 months, median age 11 months).

**Data processing**

Data was handled similarly as in surveys 1 and 2.

**Dataset 4: Wordbank MB-CDI**

**Methods**

**Participants**

**Data processing**

**Analyses**

**Age Analyses**

In Survey 1, more than 40% of parents reported that their children had produced a first word by 10 months. Because the canonically cited appearance of a first word is around 12 months (citation here), the youngest that parents could report a first word on Survey 1 was “10 months or before.” Given these results, Surveys 2 allowed parents to respond anywhere from 0 – 24 months, and Survey 3 was a free response. For Dataset 4, we examined children who only produced one word. Thus, age data for the Turk, Info, and Wordbank datasets were more varied. Taking these datasets, we graphed the cumulative probability of a child producing a first word as a function of age (Figure 1). While a child is significantly more likely to have uttered a first word by 12 months, the data suggest that some children are indeed producing words before 12 months.

* Talk a bit about over identification of babble?
* 10m? ehh



Figure 1.

**CDI Categories**

We next turned to investigate another aspect of variability in early language production, semantic category (Citation?); additionally, we explored the possibility of developmental differences in differential production for certain semantic categories. While the CDM and Mechanical Turk surveys both prompted parents to assign a CDI-Category to their children’s first words, a researcher separately validated responses, and corrected misclassified words. MB-CDI category-word mappings generated by the CDM and Mechanical Turk data processing were used for any subsequent classification of CDI categories. For all the datasets, we split the data at 12 months, and graphed the occurrences of a word belonging to that semantic category as a proportion of the total words in that dataset and age group.

* Talk about the categories that are more often represented.
* What makes these more likely to be a first word?
  + Important to babies?
  + Ease of production?
  + Input Frequency?



Figure 2

**Input Frequency Analysis**

Parental input predict first words – this needs to be done.

**Discussion**

**Acknowledgements**

**References**

Outline

* Intro
* Methods of data collection
  + CDM Survey
  + MTurk Survey
  + Childes Info Survey
  + Wordbank data
* Analyses
  + Age of onset for all four data sets
    - Data processing
    - Results and Discussion
  + CDI-Categories
    - Data processing
    - Results and Discussion
  + Input frequencies for Childes corpus
    - Data processing
    - Results and Discussion
* General Discussion