Talk, You're On Camera! Or, Comparing Naturalistic Audio and Video Recordings of Infants

Elika Bergelson[1], Andrei Amatuni[1], Shannon Dailey[1], Sharath Koorathota[3], & Shaelise Tor[4]

[1] Duke University

[2] University of Rochester

[3] Columbia University Medical Center

[4] Syracuse University

Abstract

Measurements of infants' quotidian experiences provide critical information about early development. However, the role of sampling methods in providing these measurements is rarely examined. Here we directly compare language input from hour-long video-recordings and daylong audio-recordings within the same group of 44 infants at 6 and 7 months. We find far denser noun input in video- than in audio-recordings across 12 measures of language quantity and lexical diversity, talker variability, utterance-type, and object presence. Although audio-recordings captured ~10 times more awake-time than videos, the noun input in them was only 2–4 times greater. Most notably, per unit time, videos featured more word-types and tokens, more questions but fewer declaratives, and more talkers. In contrast, >33% of videos lacked certain noun input altogether, e.g. reading and fathers' speech. While we find moderate correlations across recording-types, the most common audio-recording nouns were far more consistent across families than top video-recording nouns. Thus, hour-long videos and daylong audio-recordings provided fairly divergent pictures of the input infants hear and learn from in their daily lives. We suggest short video-recordings may inflate various language input estimates and should be used cautiously for extrapolation about common words, talkers, utterance-types, and contexts at larger timescales. If theories of language development are to be held accountable to 'facts on the ground' from observational data, greater care is needed to unpack the ramifications of sampling methods of early language input.

Talk, You're On Camera! Or, Comparing Naturalistic Audio and Video Recordings of Infants

Highlights

- We measured 44 infants' early noun input during free-form, infant-caregiver interactions in hour-long videos and daylong audio-recordings at 6 and 7 months.
- Across measures of quantity, utterance-type, object presence, and talker, nouns-per-minute were 2–4 times more frequent in videos than in audio-recordings.
- Videos had relatively more questions and fewer declaratives. The most frequent nouns across audio-recordings also occurred in most families; this was not true for videos.
- Methodological differences in naturalistic observation techniques greatly influence researchers' potential conclusions about infants' language input.

Researchers have studied development by observing infants experiencing their natural habitats for over a century (Taine, 1876; Williams, 1937). Over the past 20–30 years, written records have been increasingly supplemented with annotated audio- and video-recordings, which have described the linguistic, social, and physical landscape in which infants learn. Such data — often shared through repositories like CHILDES and Databrary — in turn provide a proxy for various "input" measures in theories of psycho-social, motor, and in particular, linguistic development (MacWhinney, 2001).

Furthermore, recent technological advances have made it feasible to collect longer, denser, and higher-quality recordings of infants' day-to-day lives, which aim to provide better approximations of infants' input and early language abilities (Bergelson & Aslin, 2017; Oller et al., 2010; B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015; VanDam et al., 2016, *inter alia*; Weisleder & Fernald, 2013). Such naturalistic data seeks to reveal what infants actually learn from as they make use of their biological endowments and environmental resources.

While cutting-edge technologies make collecting observational data ever easier, this growing toolbox increases researchers' decision load, with serious but underexplored side-effects. For instance, researchers must decide on recording modalities (e.g. audio, video,

or both), where, whom, and how long to record, and whether to capture structured or free-ranging interactions, with or without experimenters present. While any path through such decision-trees may lead to equivalent results, this is rarely tested directly. Problematically, this leads to research with theoretical conclusions built on equivalency assumptions that go unmeasured.

In recent work directly comparing observational sampling methods, Tamis-LeMonda, Kuchirko, Luo, Escobar, and Bornstein (2017) analyzed mother-infant behavior in 5-minute structured interactions and 45 minutes of free play. Home sessions were video-recorded by an experimenter and transcribed. The results showed that relative to free play, infants generally experienced more language both in word-quantity (i.e. tokens) and word-variability (i.e. types) per minute in structured interactions. They also found that language quantity across contexts correlated and that the peak five minutes of the naturalistic interaction were similar to the 5-minute structured interaction. They conclude that sampling must be matched with research question, cautioning that while brief samples may be appropriate for studying individual differences, extrapolations about overall language input from short samples must be made with care.

In contrast, work by Hart and Risley (1995) extrapolated extensively. Based on 30 hours of data per family (collected one hour per month for 2.5 years), these researchers estimated that by age four, children receiving public assistance (n=6) heard >30 million fewer words than professional-class children (n=13). While their results highlighting SES differences certainly merited (and received) follow-up (e.g. Fernald, Marchman, & Weisleder, 2013; Noble, Norman, & Farah, 2005, *inter alia*), they have also been criticized as an extreme over-extrapolation (Dudley-Marling & Lucas, 2009; Michaels, 2013).

Still other research analyzes base rates of certain linguistic phenomena to provide in-principle proof of what young children can learn from their input (Brent & Siskind, 2001; Lidz, Waxman, & Freedman, 2003; Tomasello, 2000). Here, the research question generally dictates what is deemed appropriate sampling. Problematically, for most exploratory work,

"appropriate" sampling is hard to premeditate. For instance, practically any length of adult speech, across wide-ranging recording parameters, will find function words (e.g. "of") at much higher rates than content words (e.g. "fork"). For questions concerning many aspects of infants' language input, however, it is largely unknown how methodological choices may bias our answers.

In the present work, we explore these issues, directly comparing hour-long video-recordings and daylong audio-recordings in a single sample of 44 infants, as part of a larger study on early noun learning. We annotated concrete nouns (generally objects, foods, animals, or body parts) said to infants or said loudly and clearly in their presence. We further annotated three properties previously linked with early language learning: (1) utterance-type, which provides syntactic and situational information (Brent & Siskind, 2001; DeBaryshe, 1993; Hoff & Naigles, 2002) (2) object presence (i.e. referential transparency) which clarifies whether the referent of a spoken word is visually appreciable (Bergelson & Aslin, 2017; Bergelson & Swingley, 2013; Cartmill et al., 2013; Yurovsky, Smith, & Yu, 2013), and (3) talker, which lets us quantify the range of speakers infants hear (Bergmann, Cristia, & Dupoux, 2016; Rost & McMurray, 2010).

This design sets up two overarching questions. First, do features of the noun input in one video-recorded hour predict these same quantities in an entire audio-recorded day? Second, do input quantities differ once time is standardized? If the noun input is equivalent and predictive across recording-types, then researchers can freely vary their observational data collection approach with impunity. If it is not, understanding the biases of various methods is critical to ensuring our learning theories consider the data quantity and variability available to learners day-to-day.

Thus, our main goal was to compare language input across four key properties (word quantity/diversity, utterance-type, object presence, and talker), as measured by hour-long videos and (separate) full-day audio-recordings. This seemingly methodological question has deep implications for developmental theory: we examine how sampling approaches may alter

Table 1

Infant ages at recordings and lab visits

Month	Video Recordings	Audio Recordings	In-lab visits
6 months	M=6;4, SD=3.2 days	M=6;7, SD=3.9 days	M=6;2, SD=3.7 days
7 months	M=7;2, SD=2.3 days	M=7;5, SD=3.3 days	NA

conclusions about the linguistic input that in turn drives early development.

Methods

Participants

Participants were recruited from an existing database of families from local hospitals, or who heard about the BabyLab from friends, family, and outreach. Forty-six participants enrolled; two dropped out in the early stages of the project leaving 44 infants in the final sample. All infants were full-term (40 ± 3 weeks), had no known vision or hearing problems, and heard ≥ 75% spoken English in the home. Participants were 95% white; 75% of mothers had a B.A. or higher. The families were enrolled in a yearlong study that included monthly audio- and video-recordings, as well as in-lab visits every other month. Here we report on the home recording data from the first two timepoints (6 and 7 months) of this study, for which participants were compensated $10. [1]

Procedures

Participants gave consent at an initial lab visit for the larger study through a process approved by the University of Rochester IRB. Questionnaires about various aspects of the family's and infant's background conducted during lab visits, not germane to the present

---

[1]We include only these timepoints because no infants had begun producing words themselves (which changes the input for reasons orthogonal to those examined here). Given the broader project aims, these timepoints alone had the entire daylong audio-recording annotated.

analysis, are reported elsewhere (Bergelson & Aslin, 2017; Laing & Bergelson, under review). Four recordings were collected for each infant: an audio- and video-recording at 6 and at 7 months, each on a different day. See Table 1.

Audio-video release forms were given to parents and collected after the audio- and video-recordings for the month were complete. Parents could opt to share the data with other authorized researchers and/or to have excerpts used for academic presentation. The released audio and video files can be accessed by registered researchers on Databrary.

Video-Recordings

Researchers visited infants' homes each month to video-record a typical hour of infants' lives from their own perspective. To achieve this, infants were outfitted with a hat or headband affixed with two small, lightweight Looxcie cameras (22g each). One camera was oriented slightly down and the other slightly up, to capture most of the infant's visual field (verified via Bluetooth with an iPad or iPhone during setup). A standard camcorder (Panasonic HC-V100 or Sony HDR-CX240) on a tripod was set up in a location that could best capture the infant. Parents were asked to move this camera with them if they changed rooms. After set-up, experimenters left for one hour.

Audio-Recordings

Audio-recordings captured a full day (up to 16 hours) of infants' language input. Parents were given infant vests and LENAs (LENA Foundation, Boulder, CO), small audio-recorders (<60g) that fit into a small pocket on the front of the vests. Parents were asked to put the vest and recorder on babies from when they awoke to when they went to bed (with the exceptions of naps and baths). Parents were permitted to pause the recorder at any time but were asked to keep such pauses minimal.

Data Processing

Details of our entire data processing pipeline are on our lab wiki (https://osf.io/cxwyz/wiki/home/). Videos were processed using Sony Vegas and in-house video-editing scripts. The footage was aligned in a single, multi-camera view before manual language annotation in Datavyu. Audio-recordings were initially processed by LENA proprietary software, which segments and diarizes each audio file; this output was then converted to CLAN format for further processing and manual annotation (MacWhinney & Wagner, 2010). Through in-house scripts, long periods of silence were demarcated in these CLAN files (e.g. when the audio vest was removed or during naps). The CLAN files were then used for manual language annotation.

Language Annotation

Recordings were annotated by trained researchers. The "sparse annotation" entailed marking each concrete noun heard by the child. This included words directed to or easily overheard by the child (e.g. words directed at a sibling next to the infant), but not distant or background language (e.g. background television). We operationalized "object words" as concrete, imageable nouns (e.g. shoe, arm). For each object word, we included the word (as said by the speaker, e.g. "teethies") and lemma (i.e. dictionary form, e.g. "tooth"), along with three properties: utterance-type, object presence, and talker. Utterance-type classified each object word utterance as declarative, question, imperative, reading, singing, short-phrase, or unclear. Short-phrase utterances include words in isolation and short, simple noun phrases (e.g. "the red ball" or "kitty's paw"). Object-presence was a binary measure of whether the object was present and attended to. Lastly, the word's talker was recorded, including live interlocutors and electronics: mother, brother, toy, etc.; talker classification was checked by a staff member with high familiarity with each family. We assessed intercoder reliability on a random contiguous 10% of the annotations in each file for the two categorical variables (utterance-type and object-presence). Reliability was moderate to

Table 2

Derived count measures

Measure	Derived Count
Quantity	Noun tokens, Noun types
Speaker	Nouns from Mother, Nouns from Father, Unique Speakers
Utterance Type	Nouns in Declaratives, Imperatives, Questions, Short-Phrases, Reading, or Singing
Object Presence	Nouns said when the referent was present and attended to

strong (utterance-type: 87% agreement, Cohen's κ=0.81; object-presence: 83% agreement, Cohen's κ= 0.65).

Results

Analysis Plan

Based on the coding scheme above, we derived 12 count measures from each recordings' annotations for each child (n=44), recording-type (audio, video), and month (6, 7). See Table 2. We then averaged the data from months 6 and 7 to increase the precision of our input estimates, and since we have no theoretically-motivated reason to predict input differences across this 4-week span (i.e. there are no developmental or linguistic milestones typically achieved at 6–7 months.) We also normalized the count measures by recording length; further details are below. While we initially anticipated analyzing multi-level models with fixed effects of recording-type and random subject-level effects, nearly all such models revealed highly non-normal residuals (by visual inspection and Shapiro Test), even when log-transformed; this limited interpretation across measures. Thus, we instead report a simple set of nonparametric analyses below. We used R for all data aggregation and analyses; the code that rendered this manuscript and all its contents is available on GitHub

here: https://github.com/BergelsonLab/talk_youre_on_camera.

For all recording-type comparisons, we look at whether our measures *differed* significantly (by two-tailed, paired Wilcoxon Test) and *correlated* significantly (by Kendall Rank Correlation) across the given groups. This approach lets us compare, e.g., whether the proportion of declaratives is indistinguishable in our audio- and video-recordings, independently of whether these values are correlated across recording-types. We applied Holm's *p* value adjustment for multiple comparisons (Holm, 1979), for the set of 12 Wilcoxon tests and the set of 12 Kendall Correlations.

Count Measures, Audio- vs. Video-recordings

Before assessing how our 12 measures of noun input scaled between hour-long video-recordings and daylong audio-recordings, we analyzed recording lengths. Modally, videos were an hour (62 min, M=60.79 min, SD=6.31, R=27.9–74.9 min), and audio-recordings were 16 hours (960 min, M=858.41 min, SD=119.41, R=635–960 min), the maximum capacity of the LENA device. While audio-recordings began when children awoke, we further estimated the onsets and offsets of daytime naps by removing the "silent" portions of the recordings (see Methods). This provided an estimated upper-limit on infants' awake (i.e. non-silent) time (Mode = 654 min., M = 603 min, SD=106.8, R=385.2–951 min). Our estimates comported with established norms for 6–8-month-olds in the US (**ADD REF:** Mandel et al, 2010), which are 180 minutes of daytime sleep, and 600 minutes of nighttime sleep. Infants were always awake during video-recordings (save one infant, who fell asleep before the recording-hour ended; that video was stopped at sleep onset).

To examine how the hour-long video data "scale" to day-length data descriptively, we first divided the 12 count metrics from the videos by those from the audio-recordings for each child, to derive "video-fraction" score. This showed that the video-recordings were 0.07 of the length of audio-recordings, or 0.10 of the length if only "non-silent" portions of the audio-recording are included. However, rather than a concomitant 10-fold decrease in our
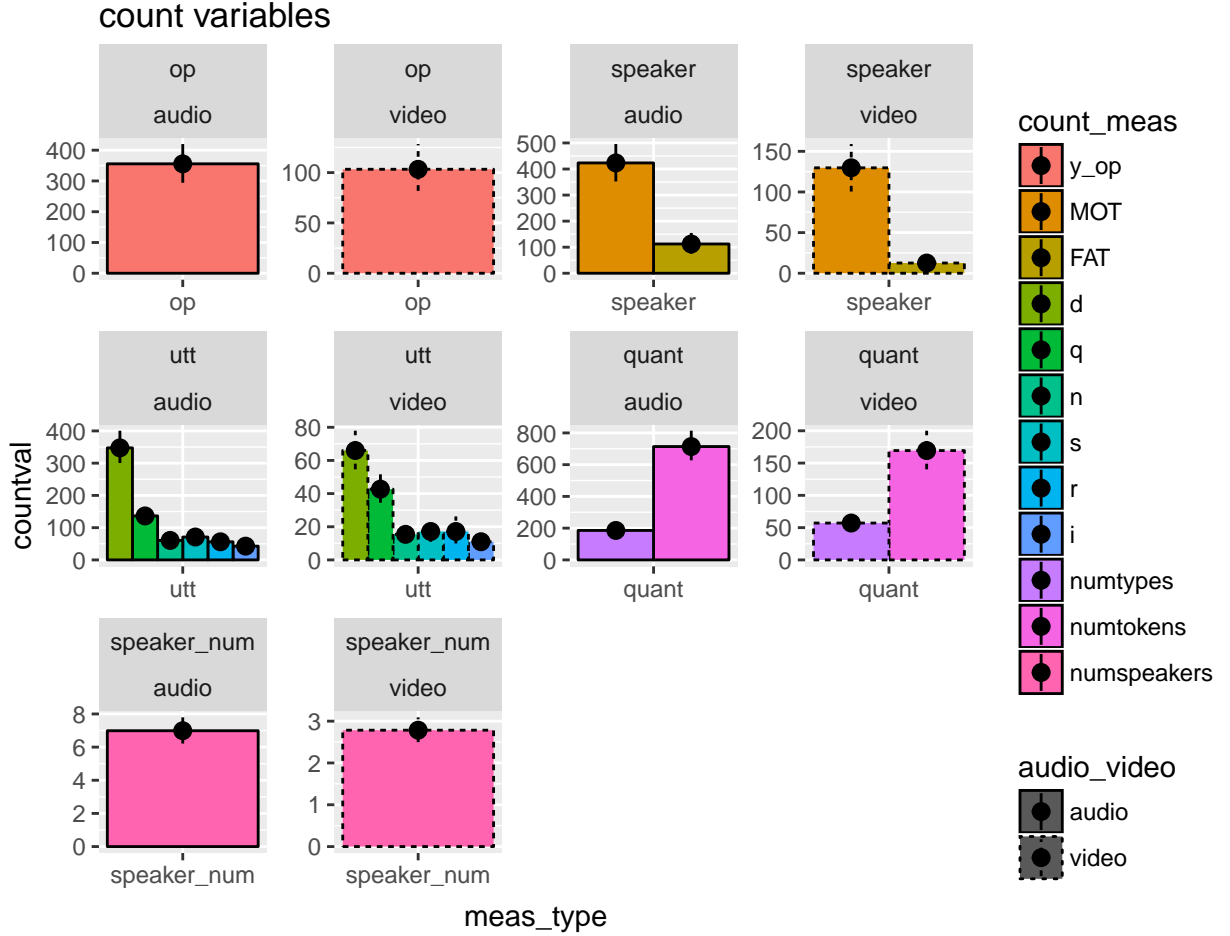
Figure 1. Raw count of variables across audio (indicated by solid borders) and video (dashed borders) recordings. Variables, in order: object presence, nouns from mother and father, utterance-type, noun-types and tokens, and count of unique speakers.

count metrics (as would be expected if the video captured a "representative" hour of the day), the fractions averaged to 0.31 across measures; see Table 4. Thus, by and large, videos had a denser concentration of nouns across our measures than did the audio-recordings. See Figure 1 for raw count data for each metric.

We computed video-fractions (rather than the reciprocal, i.e. audio/video) because there were more zero values for videos than audio-recordings (e.g. instances when children did not hear any nouns sung), rendering more undefined values. Indeed, $>1/3$ of children did not hear nouns in reading or from fathers on videos in either month. See Table 3.

Table 3

Proportion of infants with no recorded nouns from the listed speakers and utterance-types

V: Mothers	V: Fathers	V: Imperatives	A: Singing	V: Singing	A: Reading	V: Reading
0.09	0.51	0.02	0.02	0.11	0.16	0.34

Note. V indicates videos, A indicates audio-recordings. All infants heard nouns in all other categories (see Table 2).

We next normed our count values by the number of minutes in each. For example, if an infant heard 500 noun-tokens in 800 minutes of non-silent audio-recording, and 200 in 60 minutes of videos, this was normed to .62 and 3.3 noun-tokens per minute, respectively. Zero values are retained within the normed count data. [2]

With the normed data, 11/12 metrics occurred at significantly lower rates in audio-recordings than video-recordings (all adjusted-p<.05). The remaining metric, number of nouns from fathers, was statistically indistinguishable across recording types (adjusted-p>.05). Thus, overall, per unit time, infants heard less noun input across our metrics of quantity, talker, utterance-type and object presence in audio-recordings than in videos. We provide both raw and normed count data in Figure 1 and 2.

Looking next at correlations, we find that 10/12 metrics correlated in audio vs. video data; nouns per minute heard from fathers and in singing did not. The size of the correlations (i.e. Kendall's τ) was moderate (excluding the two non-significant metrics, $M =$ 0.44, 0.27 - 0.57, all adjusted-p<.05). See Table 4 and Figure 3.

Exploratory Analyses over Utterance-Types and Nouns

Lastly, we undertook two sets of highly exploratory analyses, at the utterance level and at the word level. The utterance-type analysis is based on the unanticipated observation that while declaratives and questions made up >2/3 of the input for each recording-type, the

---

[2]One infant's zero value was excluded from 'father' measures because this infant had no father.

Table 4

Video/Audio Count Measures, normed by minutes in recording (column 2) and divided without norming (column 3)

Measure	Inflation (normed)	Video-fraction Mean(SD)
Minutes	NA	0.07 (0.01)
Awake minutes	NA	0.1 (0.02)
Types	3.00	0.31 (0.13)
Tokens	2.30	0.25 (0.15)
Speakers	3.90	0.43 (0.2)
Mother	3.00	0.32 (0.22)
Father	1.10	0.13 (0.26)
Declaratives	1.90	0.19 (0.09)
Questions	3.10	0.33 (0.16)
Imperatives	2.60	0.27 (0.23)
Singing	2.30	0.65 (1.46)
Reading	2.90	1.02 (2.76)
Short phrases	2.50	0.3 (0.25)
Object presence	2.90	0.34 (0.28)

Note. If videos contained equivalent quantities of nouns, Inflation values would be 1, and Video-fractions would be .1
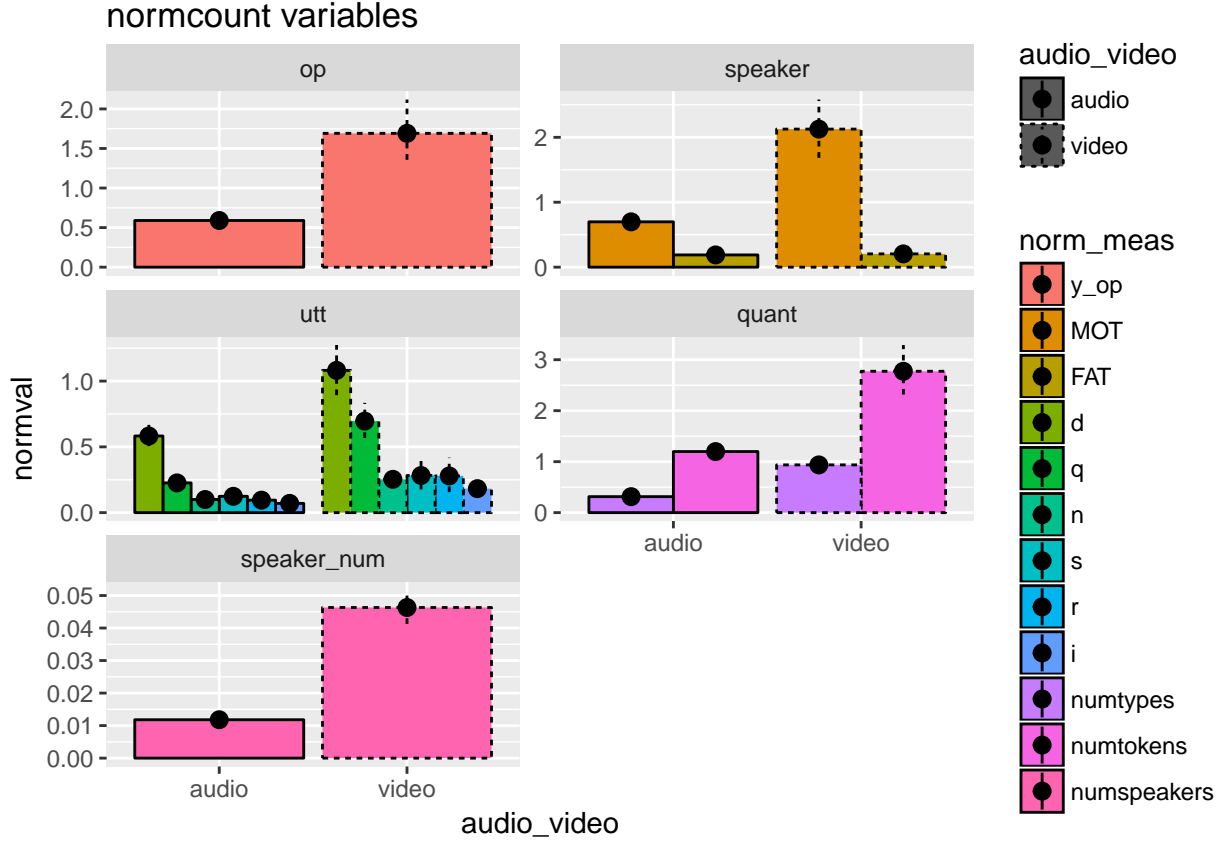
Figure 2. Normalized variable counts across audio-recordings (solid borders) and videos (dashed borders). Normed counts are calculated as the proportion of raw counts over total non-silent recording time (in minutes). Variables, in order: object presence, nouns from mother and father, utterance-type, noun-types and tokens, and count of unique speakers.

videos appeared to contain relatively more questions and fewer declaratives (See Fig 1 and Fig 2). To test this statistically, we converted the six utterance-type counts to proportions (e.g. number of nouns heard in declaratives over total noun tokens) for each recording-type. Wilcoxon tests of each utterance-type in audio- vs. video-recording (corrected for multiple comparisons as before) revealed that indeed, declaratives and questions occurred at different rates in audio- and video-recordings (both adjusted-p<.05), with audio-recordings containing relatively fewer questions (M_{video}=0.26, M_{audio}=0.19) and more declaratives than videos (M_{video}=0.40, M_{audio}=0.50). No other proportional utterance-type differences reached significance across recording-types (all adjusted-p>.05). See Figure 4.
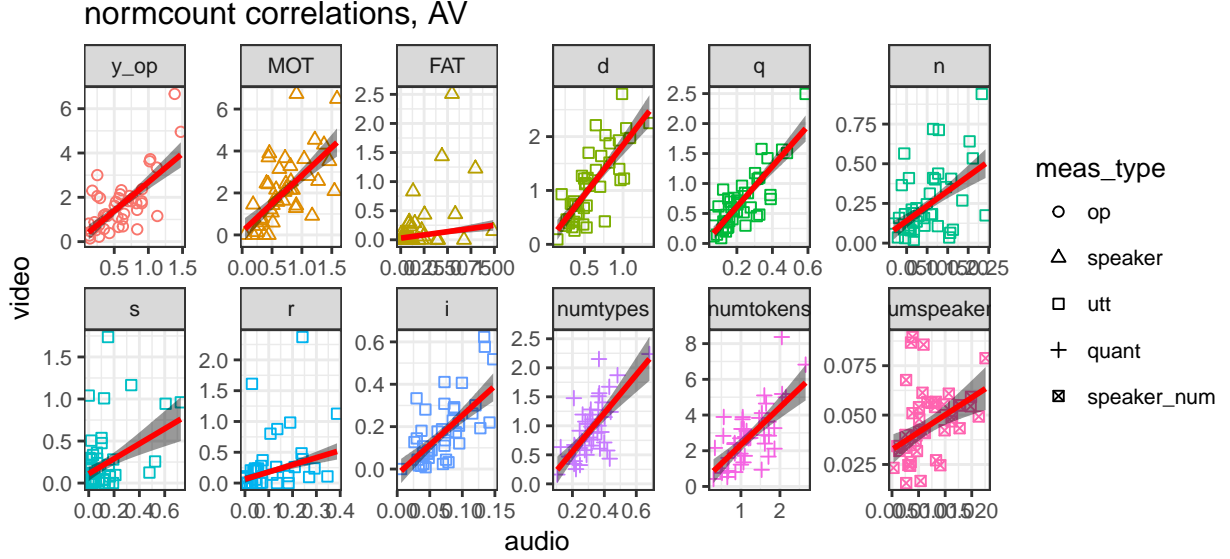
Figure 3. Normalized count correlations between audio- vs. video-recordings. Each point indicates nouns per minute of recording for each child, averaged across months 6 & 7. Variables in order: object presence, nouns from mother and father, utterance-type, noun-types and tokens, and unique speakers.

Our final analysis is at the word level; we aim to provide a first-pass characterization of whether audio- and video-recordings captured the same nouns and the same relative frequencies across words and families. The distribution of nouns in our recordings was Zipfian: of the 5801 unique object words (3137 lemmas) heard across months and recording types, only 2482 (960 lemmas) were heard more than once.

We examined the top 100 most frequent nouns from audio- and video-recordings (n=136 due to ties, n=68 without words that occurred zero times in one recording-type). Frequency across recording-types correlated significantly (Kendall's tau: 0.39, $p<.0001$) even with zero-frequency words included (Kendall's tau: 0.25, $p<.0001$; see Figure 6 and 5).

Finally, looking at just the top ten words by recording-type, we find several notable results. Firstly, four of the top ten words in each recording-type overlapped (baby, book, mouth, toes), suggesting that extremely common words are relatively conserved across recording-types. However, the top audio words were far more common across families (see
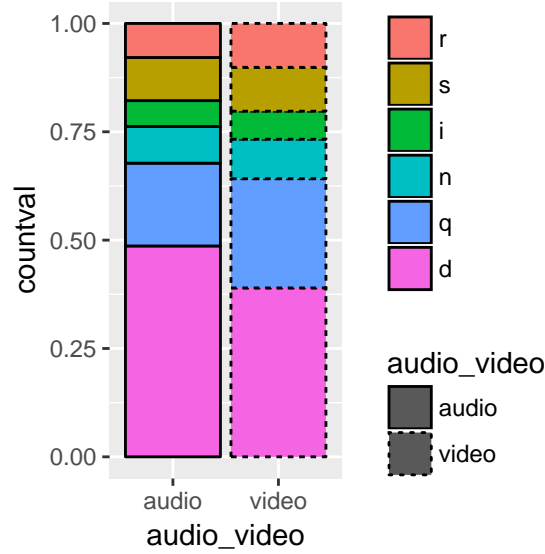
Figure 4. Utterance-type proportions across audio-recordings (solid borders) and video-recordings (dashed borders). Utterance-types top to bottom are r: reading, s: singing, i: imperative, n: short phrase, q: questions, and d: declaratives.

Figure 7). Indeed, the ten most frequent nouns in audio-recordings were heard by 37-44 (M = 42.30 (2.63)) of the 44 families, while those in video-recordings were heard by 23-42 (M = 31 (6.27). Finally, the top audio words were ~3 times as common as the top video words (M_{audio}=761.80(114.75), M_{video}=232.80(91.38)), again underscoring the higher density of nouns in video-recordings (which were 1/10 the length of audio-recordings on average). Taken together, this exploratory analysis suggests that daylong audio-recordings may render more stable estimates of pervasively common words across families than do video-recordings.

Discussion

Our results can be distilled to three key findings. First, infants heard relatively more nouns in the video-recordings than in the audio-recordings. Per minute, infants heard ~2–4 times more noun input across our quantity, speaker, utterance-type, and object-presence metrics when they and their caretakers were video-recorded for an hour versus audio-recorded for a day. Second, while our metrics generally correlated across audio- and
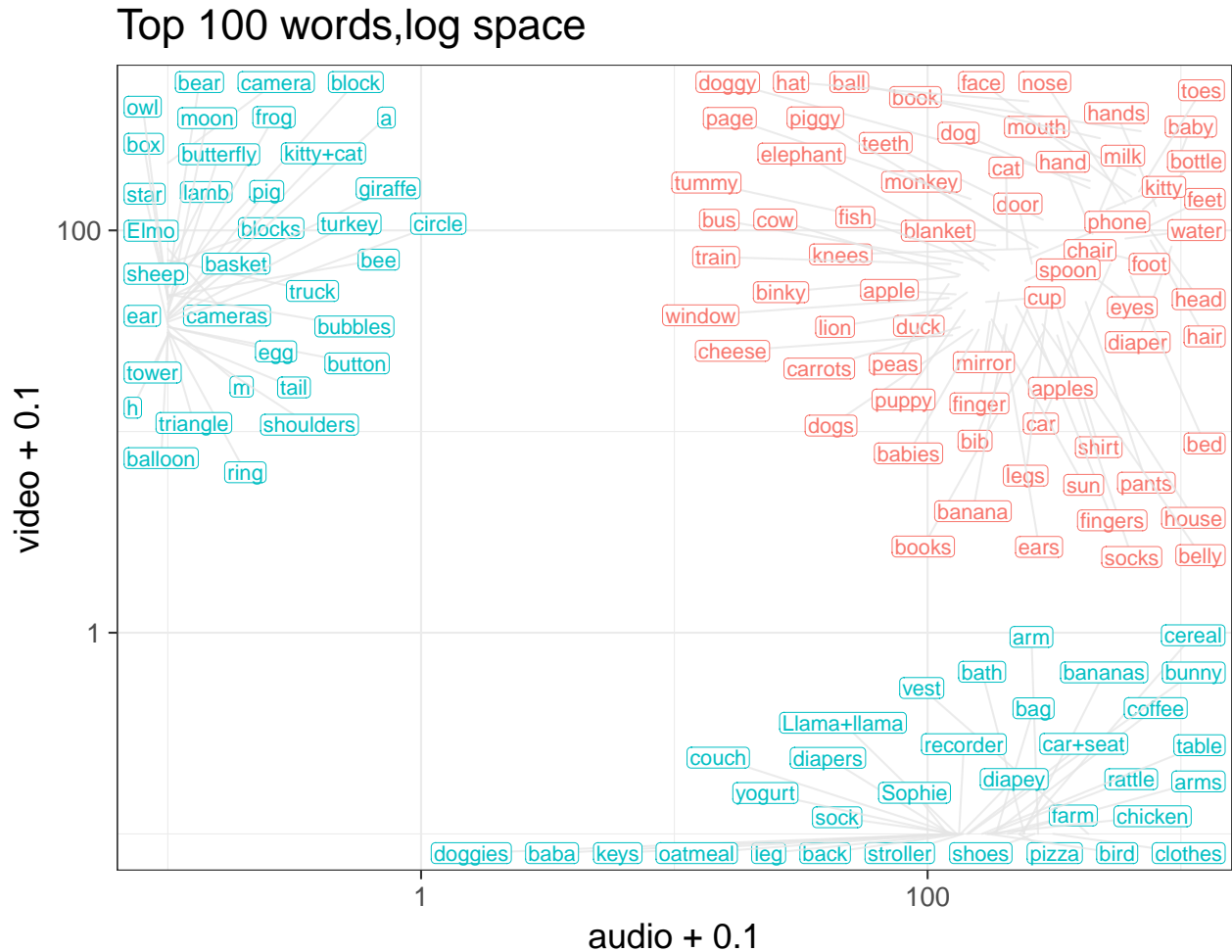
Figure 5. Log-scaled counts of the top 100 words in audio- and video-recordings. Each node represents the averaged count, across all participants in both months, of a unique noun-token. Words that occurred zero times in one recording-type are not shown.

video-recordings, the relative rates of the most prevalent utterance-types and the rates of unattested data for certain metrics varied across them. Finally, while the highest frequency words across recording types largely overlapped and correlated (and exhibited Zipfian frequency distributions), top words from the daylong audio-recording appear to better represent the noun input across families.
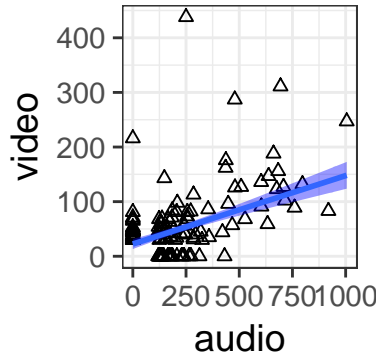
Figure 6. Correlations of the top 100 words in audio- vs. video-recordings. Each node represents the count of a unique word that occurred in both audio- and video-recordings, averaged across all participants in both months.

Noun Quantity and Lexical Diversity

The pattern across recording-types suggests to us that parents behaved naturally during recordings, but that "natural" behavior differed by recording context. This is consistent with a point made by D. Suskind et al. (2013) regarding an intervention: "sustaining increased talk for a 10-hr recording day is much less likely than being on best behavior during [a] 1-hr videotaped session. . . " While their work aimed to encourage caretakers to talk more, the point stands for our goals of observing infants' typical input. We add to their suggestion that a shorter video-recording itself may elicit certain kinds of interactions, separate from deliberate intent or lack thereof on caretakers' part.

Indeed, the kinds of everyday interactions we captured in daylong audio-recordings (family members rushing to get out the door or get meals on the table, sibling quibbles, etc.) tended to "feel" more natural. Families likely simply found it easier to go about their day freely with infants in a special vest than with a camera on their head and a camcorder in the corner. Lending some support to the idea that equipment prominence matters, both "hat" and "camera" are in the top 10 words from video-recordings each month; no analogous nouns (e.g. vest, recorder) topped the frequency rankings in our audio-recordings (see Figures 7 and 5).
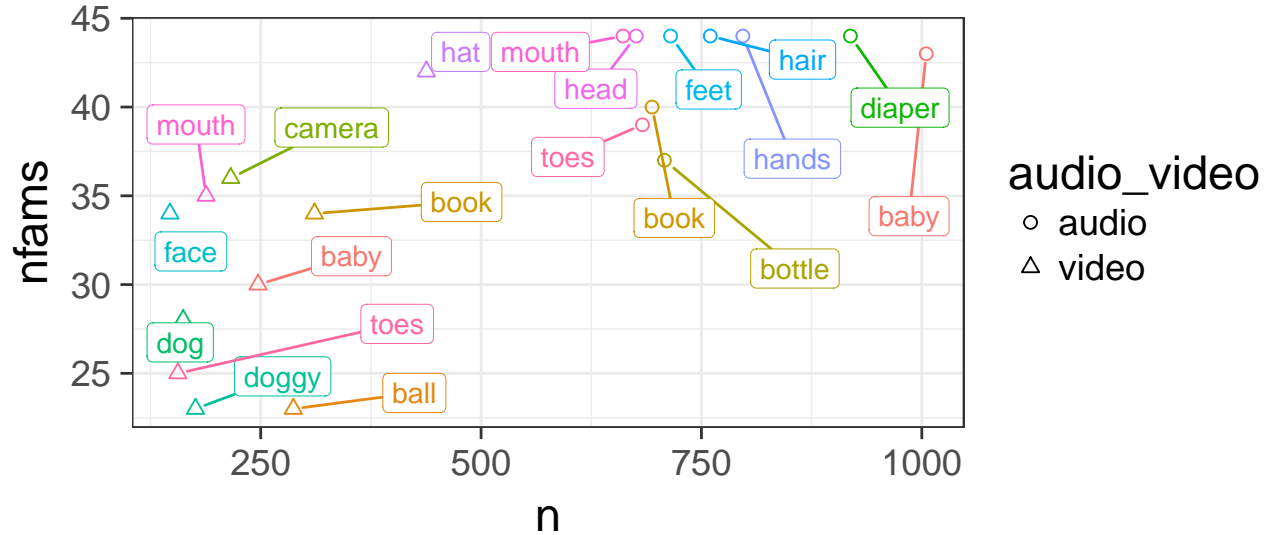
Figure 7. Top 10 words by recording type. Each node represents the frequency count of a top audio or video word over both months (x-axis) and count of participants who were exposed to the word in either or both months (y-axis).

Our comparison across recording-types highlighted many differences across our noun input measures, even with family and age held constant. The quantity metrics provide a conceptual replication and extension of Tamis-LeMonda et al. (2017). Despite numerous methodological differences (length and type of recordings compared, experimenter presence, infant age, word class analyzed), both studies found that parent talk per unit time was significantly higher in shorter recordings. While the difference they find is less extreme numerical (roughly 1.5–2 times the number of types and tokens in the longer vs. shorter recording compared to our 2–3-fold difference), this general pattern appears robust across our very different sampling methods. Taken together, these results converge in suggesting that shorter recordings elicit denser caregiver talk.

For certain research questions, these differences in volubility and lexical diversity may not matter. For instance, for studies aiming to compare relative rates of word use and object interactions during a concentrated exposure and test phase in the lab, higher volubility and lexical diversity in shorter recordings may be less relevant. In contrast, research aiming to quantify language input across populations with varying demographic, social, and cultural

properties may need to be particularly sensitive to cross-sample comparison in word quantity, as a function of sampling parameters (cf. Bergelson et al., under review; Cristia, Dupoux, Gurven, & Stieglitz, n.d.; Shneidman & Goldin-Meadow, 2012).

Object Presence

We also found more object presence in videos than in audio-recordings. This may be because the video-recordings truly had more object presence (i.e. infants mostly stayed in 1–2 rooms, interacting with caregivers and objects at hand). Alternatively, it may be the case that there are more ambiguous cases of "object presence" in audio-recordings than video-recordings, which may have contributed to "not present" annotations at higher rates. Indeed, although object presence did correlate significantly across recording-types (0.40), inter-rater reliability was higher for videos than audio-recordings for this measure (audio: 78.68% agreement, Cohen's κ=0.57; video: 87.21% agreement, Cohen's κ=0.73). Our interpretation is that both factors are likely at play, i.e. that the object-presence difference we find reflects a true difference between situations that arise during daylong audio vs. hour-long video-recordings, and that there is more noise in the estimate of object-presence when visual information is unavailable. Given that object presence and the related ideas of referential transparency and contingent talk have been linked with early word learning and language development ((Bergelson & Aslin, 2017); (Yurovsky et al., 2013); (McGillion, Pine, Herbert, & Matthews, 2017); (Cartmill et al., 2013)), this property in particular merits follow-up. That is, a more thorough understanding of the situations and contexts that elicit and promote contingent, referentially-transparent caretaker talk (around objects or otherwise) may be a fruitful avenue for further work.

Talker Variability

Our talker-based results revealed that infants heard nouns from more talkers per minute in videos than in audio-recordings, though in raw numbers infants heard roughly double the speakers over the course of a day as they heard in one video-recorded hour (see

Figure **??**. While we considered noun input from all sources (human, electronic, etc.), the quantity of talkers was largely swamped by the preponderance of the input (~65%) that came from infants' mothers; this quantity was also greater in videos than audio-recordings. The proportion of input from fathers did not vary by recording-type, though over half of videos did not include noun input from fathers at all. This is largely due to the demographics of our sample: we video-recorded during regular business hours during weekdays, a time when the fathers in our sample were largely at work. In contrast, audio-recordings could happen on any day of the week, at parents' leisure. Given that fathers and mothers make different contributions to early language development (Pancsofar & Vernon-Feagans, 2006), this is a clear example of a consequence of methodological choices: to better understand the role of input from fathers, it is critical to sample with work schedules in mind.

Questions concerning talker variability in the input are also relevant for recent in-lab studies: while infants at the same age tested here looked equivalently to a named target image when words were produced by a new person or their mother (Bergelson & Swingley, 2017), slightly older infants show a word-learning advantage when they heard multiple talkers naming new objects (Rost & McMurray, 2010). Other recent work has highlighted that certain phonetic discriminations are differentially affected when talker variability is considered (Bergmann et al., 2016). One general goal of such research is to test the proposed mechanisms by which infants generalize over or utilize talker variability during early learning. To this end, benchmarking how much talker variability is in infants' quotidian experience is desirable, even if only to say that a given model is unaltered by deviations within a given range. The present results find that these estimates are inflated in hour-long video-recordings relative to daylong audio-recordings.

Utterance-Types

Turning to our utterance-type metrics, we found more nouns in every utterance-type in videos than in audio-recordings, per unit time. The utterance-types we annotated were a mix

of largely syntactic constructions (declaratives, questions, imperatives, short phrases) and more situationally-defined utterance-types (reading, singing).[3] While its not particularly surprising that rates of reading or singing might vary across recording-types, we did not anticipate differences in relative rates of declaratives and questions, the utterance-types that made up most of the input. We found that while questions and declaratives made up the majority of the input for each recording-type at each month, videos had relatively more questions and fewer declaratives. This is key example of methodological choices potentially influencing language acquisition theories: base rates of interrogatives taken from videos would inflate estimates of auxiliary verbs in the early input. Indeed, previous work has noted that published studies vary in whether they find links between questions (yes/no and wh-) in the input and children's early productions, with developmental level of the child invoked to explain differences across studies (Barnes, Gutfreund, Satterly, & Wells, 1983; see discussion in Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002). Here we add the possibility that recording-type too may contribute to the base-rates of questions in the input, even with age kept constant.

The reading data also highlight interesting differences across recording-types. Twice as many infants were not read to in videos, compared to audio-recordings (see Table 3). However, there were still significantly more nouns-per-minute heard in reading in videos, and nouns in reading across recording-types also correlated significantly. This pattern highlights the differences in answers about "reading in the input" that can be garnered by even slightly different formulations of the relevant question. Given the clear benefits in input richness and language abilities linked to reading in infancy found in previous research (DeBaryshe, 1993; Montag, Jones, & Smith, 2015), and the likelihood that reading occurs at higher rates at certain parts of the day (a research question in its own right!), we suggest here that considering reading input on the daylong scale may give a more accurate picture of its presence in infants' lives.

---

[3]When necessary, we used prosody to disambiguate, e.g. "Get your blocks?" was coded as a question

Top Words

Our interpretation of the present results is that findings based on relatively short video-recordings overestimate young infants' typical noun input, and that extrapolation based on daylong audio-recordings likely better represents infants' daily lives. This underscores our third main finding: the conclusions one would draw about which words are most common in young infants' language input differ in their robustness across families by recording-type. That is, the top audio words were all heard by $\geq 75\%$ of the families we recorded; only one of the top 10 video words ("hat") was this common across families, and was a clear artifact of our hat-mounted video-cameras. This result may be meaningful in several ways. From a high-level viewpoint, corpora of child language input offer our best proxies for what infants learn from: our "top words" analysis suggests that one would be led to believe that the input is far more heterogeneous across children than it really is based on hour-long video-recordings alone. At a lower level, word frequency and prevalence across families are often used to select stimuli for in-lab study; relying on estimates from shorter, less representative recordings may stymie the words we chose to study in the lab. Thus, understanding how cross-family noun-input stability scales with recording-length may prove critical for future research; the word-level results above are an initial exploration in understanding this dimension of naturalistic observational data.

Limitations and Conclusions

Given the technical limitation that currently available infant-friendly video-recorders have a shorter battery life than audio-recorders, we cannot conclusively separate the effects of modality and length. That is, had we only audio-recorded for an hour or recorded video all day, we may have obtained equivalent results across recording modalities. Such a comparison awaits technological progress. A further limitation is the likely influence of self-selection into the study: many parents are unwilling to invite researchers to record their infants' interactions. Relatedly, our convenience sample does not reflect the broader

demographics of the US (let alone other cultures or populations), and as such this work merits extension to other populations before conclusive generalizations about sampling methodology can be made (cf. Bergelson et al., under review).

Understanding what infants learn from is a key part of understanding what and how they learn at all. Here we have taken first steps in understanding how two different data collection approaches may influence our conclusions about early linguistic input. We find that even naturalistic observer-free video-recordings appear to inflate language input relative to daylong recordings, in ways that influence syntactic constructions, word-specific experiences, talker-variability, and the sheer quantity and diversity of nouns infants hear. Work from the preceding decades suggests all of these factors matter for early learning. Yet without knowing how our sampling methods may be limiting us in principle, we necessarily limit our ability to adequately model infant language acquisition. The present work charted datapoints within this largely underspecified space, probing how robust linguistically-relevant measures are across two naturalistic sampling methods of infants' everyday experiences.

References

Barnes, S., Gutfreund, M., Satterly, D., & Wells, G. (1983). Characteristics of adult speech which predict children's language development. *Journal of Child Language*, *10*(1), 65?84. doi:10.1017/S0305000900005146

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bergelson, E., & Swingley, D. (2017). Young infants' word comprehension given an unfamiliar talker or altered pronunciations. *Child Development.*

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A., & Amatuni, A. (under review). What do north american babies hear? A large-scale cross-corpus analysis.

Bergmann, C., Cristia, A., & Dupoux, E. (2016). Discriminability of sound contrasts in the face of speaker variation quantified. In *Proceedings of the 38th annual meeting of the cognitive science society* (Vol. 510).

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (n.d.). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development.* doi:10.1111/cdev.12974

DeBaryshe, B. D. (1993). Joint picture-book reading correlates of early oral language skill. *Journal of Child Language*, *20*(2), 455–461.

Dudley-Marling, C., & Lucas, K. (2009). Pathologizing the language and culture of poor

children. *Language Arts*, *86*(5), 362–370.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language
processing skill and vocabulary are evident at 18 months. *Developmental Science*,
16(2), 234–248.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young
american children.* Paul H Brookes Publishing.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child
Development*, *73*(2), 418–433.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian
Journal of Statistics*, *6*(2), 65–70. Retrieved from
http://www.jstor.org/stable/4615733

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and
child syntax. *Cognitive Psychology*, *45*(3), 337–374.

Laing, C., & Bergelson, E. (under review). The effect of mothers' work schedule on
17-month-olds' productive vocabulary.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't
have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*,
89(3), 295–303.

MacWhinney, B. (2001). Emergentist approaches to language. *TYPOLOGICAL STUDIES
IN LANGUAGE*, *45*, 449–470.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The clan
software and the talkbank data repository. *Gesprachsforschung: Online-Zeitschrift
Zur Verbalen Interaktion*, *11*, 154.

McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled
trial to test the effect of promoting caregiver contingent talk on language
development in infants from diverse socioeconomic status backgrounds. *Journal of
Child Psychology and Psychiatry.*

Michaels, S. (2013). Commentary: Déjà vu all over again: What's wrong with hart & risley and a" linguistic deficit" framework in early childhood education? *Learning Landscapes*, *7*(1), 23–41.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496.

Noble, K. G., Norman, M. F., & Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental Science*, *8*(1), 74–87.

Oller, D. K., Niyogi, P., Gray, S., Richards, J., Gilkerson, J., Xu, D., . . . Warren, S. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354–13359.

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology*, *27*(6), 571–587.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Shneidman, L., & Goldin-Meadow, S. (2012). Language input and acquisition in a mayan village: How important is directed speech?, *15*, 659–73.

Suskind, D., Leffel, K. R., Hernandez, M. W., Sapolich, S. G., Suskind, E., Kirkham, E., & Meehan, P. (2013). An exploratory study of "quantitative linguistic feedback": Effect of lena feedback on adult language production. *Communication Disorders Quarterly*, *34*(4), 199–209. doi:10.1177/1525740112473146

Taine, H. (1876). Note sur l'acquisition du langage chez les enfants et dans l'espèce humaine. *Revue Philosophique de La France et de L'Etranger*, 5–23.

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science.*

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*(3), 209–253.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142). Thieme Medical Publishers.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24*(11), 2143–2152.

Williams, H. M. (1937). An analytical study of language achievement in preschool children. *University of Iowa Studies in Child Welfare, 13*, 9–18.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*(6), 959–966.