Day by Day, Hour by Hour: Naturalistic Language Input to Infants

Elika Bergelson[1,2], Andrei Amatuni[1,2], Shannon Dailey[1,2], Sharath Koorathota[2,3], & Shaelise Tor[2,4]

[1] Duke University

[2] University of Rochester

[3] Columbia University Medical Center

[4] Syracuse University

Author Note

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 compared 12 measures of language quantity and lexical diversity, talker variability, utterance-type, and object presence, finding moderate correlations across recording-types. However, video-recordings generally featured far denser noun input across these measures compared to the daylong audio-recordings, more akin to 'peak' audio hours (though not as high in talkers and word-types). Although audio-recordings captured ~10 times more awake-time than videos, the noun input in them was only 2–4 times greater. Notably, whether we compared videos to daylong audio-recordings or peak audio times, videos featured relatively fewer declaratives and more questions; furthermore, the most common video-recorded nouns were less consistent across families than the top audio-recording nouns were. Thus, hour-long videos and daylong audio-recordings revealed fairly divergent pictures of the language infants hear and learn from in their daily lives. We suggest short video-recordings provide a dense and somewhat different sample of infants' language experiences, rather than a typical one, 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.

Keywords: language acquisition, naturalistic observational data, infants, early home environment, language input, cognitive development

Word count: 3979

Day by Day, Hour by Hour: Naturalistic Language Input to Infants

Highlights

- We measured 44 infants' early noun input during free-form interactions in hour-long videos and daylong audio-recordings; sampling approach shifted potential conclusions about home language environment.

- Across quantity, utterance-type, object presence, and talker measures, nouns-per-minute were 2–4 times more frequent in video- than audio-recordings; videos were similar to *peak* audio hours.

- Nouns in videos occurred relatively more often in questions and less often in declaratives than they did in daylong or peak audio-recording hours.

- The most frequent nouns in daylong and peak audio-recording hours highly overlapped in identity and across families; this was less true for top video nouns.

Researchers have long studied development by observing infants in their natural habitats (Taine, 1876; Williams, 1937). Over the preceding decades, written records have been increasingly supplemented with audio- and video-recordings, depicting infants' linguistic, social, and physical landscape. Such data — often shared through repositories like CHILDES and Databrary — provide a proxy for various "input" measures in theories of social, motor, and particularly, *linguistic* development (MacWhinney, 2001).

Furthermore, recent technological advances have harnessed longer and denser recordings to study infants' input and language skills (Bergelson & Aslin, 2017; Oller et al., 2010; B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015; VanDam et al., 2016; Weisleder & Fernald, 2013, *inter alia*). Such naturalistic approaches aim to characterize children's actual learning environment to better understand language acquisition.

However, wider-ranging technology creates more decision-points. Researchers must decide on recording modalities (e.g. audio, video), where, whom, and how long to record, and whether to capture structured or free-ranging interactions, with or without experimenters

present. The equivalence of these decisions is rarely tested. Problematically, this leads to research with *theoretical* conclusions built on unmeasured *methodological* assumptions.

Recently, Tamis-LeMonda, Kuchirko, Luo, Escobar, and Bornstein (2017) directly compared sampling methods by analyzing mother-infant behavior in 5-minute structured interactions and 45min. of free play. They found that while language quantity across contexts correlated, infants experienced more words per minute in structured interactions than in free play. They conclude that sampling must match the research question, cautioning that extrapolations from short samples merit extra care.

In contrast, work by Hart and Risley (1995) extrapolated extensively. Based on 30 recorded hours per family (collected over 2.5 years), they 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 merited and received follow-up (e.g. Fernald, Marchman, & Weisleder, 2013; Noble, Norman, & Farah, 2005, *inter alia*), they have also been criticized as extreme over-extrapolation (Dudley-Marling & Lucas, 2009; Michaels, 2013).

Still other research analyzes base rates of certain linguistic phenomena through child corpora (Brent & Siskind, 2001; Lidz, Waxman, & Freedman, 2003; Tomasello, 2000). Unfortunately, predetermining "appropriate" sampling for such base rates is difficult. For instance, practically any length of adult speech will find function words (e.g. "of") at much higher rates than content words (e.g. "fork"). For many questions, however, potential sampling bias is unknown, leaving practical constraints to guide sampling parameters.

We explore sampling 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 said to infants, focusing on nouns given their prevalence in early lexicons (Dale & Fenson, 1996). We further annotated three properties previously linked with early language learning: utterance-type, which provides syntactic/situational information (Brent & Siskind, 2001; DeBaryshe, 1993; Hoff & Naigles, 2002), object presence (i.e. referential transparency), which tags whether spoken words'

referents are present and attended to (Bergelson & Aslin, 2017; Bergelson & Swingley, 2013; Cartmill et al., 2013; Yurovsky, Smith, & Yu, 2013), and talker, which measures talkers' quantity and prevalence (Bergmann, Cristia, & Dupoux, 2016; Rost & McMurray, 2010).

This design sets up two overarching questions. First, does noun input in one video-recorded hour predict noun input in an audio-recorded day? Second, do input quantities differ once time is normalized? If the input is equivalent and predictive across recording-types, then observational data-collection approaches can vary with impunity. If not, understanding methodological biases can help learning theories incorporate appropriate bounds on data quantity and variability.

Thus, we examine home recordings across four key properties of language input: word quantity, utterance-type, object presence, and talker. This seemingly methodological investigation has deep implications for developmental theory: we examine how sampling may alter conclusions about the linguistic input driving early development.

Methods

Participants

Infants were recruited from a database of local families. Forty-six participants enrolled; two dropped out leaving 44 in the final sample. All were full-term (40±3 weeks), had normal vision and hearing, and heard $\geq 75\%$ spoken English. Participants were 95% white; 75% of mothers had $\geq$B.A. Families were enrolled in a yearlong study that included monthly audio- and video-recordings, as well as in-lab visits every other month. See Table 1 for age details. Here we report on home recording data from the first two timepoints (6 and 7 months) of this study, for which participants received $10.[1]

---

[1]We used these timepoints because infants did not yet produce words themselves (which changes the input). Given the broader project aims, these timepoints alone had the entire daylong audio-recording annotated.

Procedures

Participants gave consent at an initial lab visit for the larger study through a University IRB-approved process. Questionnaires concerning participant background, not germane here, are reported elsewhere (Bergelson & Aslin, 2017; Laing & Bergelson, under review). Four recordings are analyzed for each infant: an audio- and video-recording at 6 and 7 months, each on different days[2]. See Table 1. Recordings that parents approved for sharing with researchers are on Databrary.

Video- and Audio-Recordings

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

Audio-recordings captured up to 16 hours of infants' input. Parents were given small audio-recorders (<60g) called LENAs (LENA Foundation, Boulder, CO), along with vests with LENA-sized chest pockets. Parents were asked to put the vest and recorder on babies from when they awoke to when they went to bed (excepting naps and baths). Parents were permitted to pause the recorder anytime but were asked to minimize such pauses.

Data Processing

Details of the entire data-processing pipeline are on OSF (https://osf.io/cxwyz/wiki/home/). Videos were processed using Vegas and in-house scripts. Footage was aligned in a single, multi-camera view before manual language annotation in

---

[2]One video is missing due to technical error.

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 (MacWhinney & Wagner, 2010). After in-house scripts marked long periods of silence (e.g. naptimes), these files were used for manual language annotation.

Modally, videos were an hour (62min., M=60.79min., SD=6.31, R=27.9–74.9min.), and audio-recordings were 16hrs. (960min., M=858.41min., SD=119.41, R=635–960min.), LENA's maximum capacity. Removing the long silences from audio-recording left ~10hrs. of audio (Mode=654min., M=603min., SD=106.8, R=385.2–951min.), consistent with established wakeful daytime norms for 6–8-month-olds in the U.S. (Mindell, Sadeh, Wiegand, How, & Goh, 2010). All infants were awake for video-recording except one, whose video annotation ended at sleep onset.

Language Annotation

Trained researchers annotated each recording. This entailed demarcating each concrete noun directed to or said loudly and clearly near the child (e.g. at adjacent siblings), but not distant language (e.g. background television). "Object words" were operationalized as concrete, imageable nouns (e.g. shoe, arm). Each annotation noted the noun and lemma (e.g. teethies, tooth), along with *utterance-type*, *object-presence*, and *talker*. *Utterance-type* classified each noun's utterance as declarative, question, imperative, reading, singing, short-phrase, or unclear. (Short-phrases included isolated words and <3-word noun-phrases, e.g. "the red ball" or "kitty's paw".) *Object-presence* coded whether objects were present and attended to (yes/no) based on linguistic context (e.g. "here's your spoon!" was coded "yes"); in videos, visual context was also used. Lastly, *talker* tagged live interlocutors and electronics, checked by staff highly familiar with each family. We assessed intercoder reliability on a random contiguous 10% of annotations in each recording for the two categorical variables (utterance-type and object-presence). Reliability was moderate 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 measures from each recording for each child (n=44), recording-type (audio, video), and month (6, 7). See Table 2. We averaged across months to increase precision and because we lacked theoretically-motivated reasons to predict cross-month differences (i.e. no developmental or linguistic milestones are typically achieved at 6–7mo.) Unfortunately, multi-level models were not viable due to highly skewed residuals (by Shapiro-Wilk Test) even when log-transformed, limiting cross-measure interpretation. Instead we report a simple set of nonparametric analyses, conducted in R. The code that rendered this manuscript is on Github.[3]

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 time-normalized counts of declarative nouns are indistinguishable in audio- and video-recordings, independently of whether they are correlated. We applied Holm's p-value adjustment for multiple comparisons (Holm, 1979) for each set of Wilcoxon tests and Kendall Correlations.

Count Measure Analysis

To examine how the hour-long video data scale to day-length data descriptively, we first divided the 12 count measures from the videos by those from the audio-recordings for each child to derive "video-fraction" scores (video/audio). We opted for video-fractions (rather than audio/video) to minimize undefined values (e.g. 34% of children heard no nouns in reading utterances in their video-recordings; see Table 3). This analysis showed that the video-recordings were 0.07 of the length of audio-recordings, or 0.10 with audio-recording silences removed. However, rather than a concomitant 10-fold decrease in the count

---

[3]Please contact corresponding author for access before publication.

measures (as would be expected if videos captured a representative hour), the fractions averaged to 0.31; see Table 4. Thus, by and large, videos had a denser concentration of nouns across measures than audio-recordings did. See Figure 1.

We next normed our counts by recording minutes. For example, if an infant heard 500 noun-tokens in 800 non-silent audio-recording minutes and 200 in 60 video-minutes, this was normed to .62 and 3.3 noun-tokens/minute, respectively; zero values were retained.[4]

Measuring correlations across recording types, we found that 10/12 metrics correlated in audio vs. video data; nouns per minute heard from fathers and in singing did not. Correlation magnitude (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.[5]

We next compared the rates of each measure in three ways. First, we used the normed data, looking at counts per minute. 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, nouns from fathers, was statistically indistinguishable across recording types (adjusted-p>.05). Thus, overall, per unit time, infants heard less noun input in audio-recordings than in videos, across metrics of quantity, talker, utterance-type and object presence (see Figure 1 and 2).

Next, we compared two different hour-long subsets from the daylong audio recording for comparison with the video-recorded hour, collectively referred to as "peak" audio times. The *top* hour was the hour in which infants heard the most nouns. Complementarily, under the logic that parents scheduled video-recordings to optimize infant alertness, we extracted that *same* hour in the daylong audio; e.g., if the video recording visit was scheduled from 2:00pm–3:00pm, we used 2:00pm–3:00pm from that child's daylong audio recording that month. Our 12 measures were then computed in both the *top* and *same* audio hours. These

---

[4]One infant's data was excluded from 'father' measures; this infant had no father at home.

[5]The same pattern emerged with raw counts, except nouns from fathers also correlated significantly (adjusted-*p*<.05).

hours only overlapped in 15/88 recordings (17%).[6]

The results in video and same audio hours were indistinguishable for 8/12 measures; the remaining 4 occurred at significantly *higher* rates in the same audio hour (all adjusted-$p<.05$): number of speakers, noun types, nouns from fathers, and nouns in declaratives. Similarly, 7/12 occurred at significantly higher rates in top audio hour than in videos (all adjusted-$p<.05$); these included those from the same audio comparison along with noun tokens, nouns in imperatives and nouns in short phrases. Taken together, the videos presented a somewhat different language input profile than the peak audio hours of the day: videos featured less input for some quantity, talker, and utterance-type measures, but were statistically indistinguishable in object presence, input from mothers, and input in other utterance-types. This same qualitative pattern held when looking at the rate of "zero" values for the peak audio hours, relative to videos and daylong audio-recordings (see Table 3).

Exploratory Analyses

Lastly, we undertook two sets of highly exploratory analyses at the utterance and word levels. The first was based on the unanticipated result that while declaratives and questions made up >2/3 of the input for each recording-type, the videos appeared to contain relatively more questions and fewer declaratives (See Fig. 1 and 2). To test this statistically, we converted the six utterance-type counts to proportions (e.g. declarative nouns/total nouns). Wilcoxon tests of each utterance-type in audio- vs. video-recording (corrected for multiple comparisons) revealed that indeed, declaratives and questions occurred at different rates across recording-types, with audio-recordings containing relatively fewer questions ($M_{video}=0.26$, $M_{audio}=0.19$, $M_{same\ audio}=0.21$, $M_{top\ audio}=0.17$) and more declaratives than videos ($M_{video}=0.40$, $M_{audio}=0.50$, $M_{same\ audio}=0.49$, $M_{top\ audio}=0.47$; each video vs. audio comparison adjusted-$p<.05$). No other proportional utterance-type differences reached

---

[6]In 3 cases, the video-recording time (i.e. 'same' time) preceded the beginning of the daylong audio-recording (by 5, 30, or 90 minutes); in those cases the first hour of the recording was used. This created two further cases of 'top' and 'same' overlap.

significance across recording-types (all adjusted-p>.05). See Figure 4.

At the word level, we aimed to characterize whether audio- and video-recordings captured similar nouns at similar relative frequencies across words and families. Nouns' frequency distribution was Zipfian: of the 5801 unique object words (3137 lemmas), only 2482 (960 lemmas) occurred >1 time.

We examined the 100 most frequent nouns from audio- and video-recordings (n=136 due to ties, n=68 excluding words that never occurred in one recording-type). Frequency across recording-types correlated significantly (Kendall's τ: 0.39, p<.0001) even with zero-frequency words included (Kendall's τ: 0.25, p<.0001; see Figure 5 and 6).[7]

Lastly, we analyzed the top ten nouns within videos, daylong-audios, and both peak audio hours. Four of the top ten words in each time sample overlapped (baby, book, mouth, toes), suggesting that extremely common nouns are relatively well-conserved. Moreover, all but one word in the top 10 was identical for all 3 audio-based time-slices, while 5 of the top video words were unique to video recordings (see Figure 7).

The top 10 words within each time sample also varied in how common they were across the 44 families: top words from daylong audio occurred in 96% of families (M=42.30(2.63); those in video-recordings were heard by 70% (M=31(6.27)). Nouns in peak audio hours patterned in between (top hour: 88% (M=38.70(2.83); same hour: 78%, M=34.20(4.71)).

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)), further underscoring the higher density of nouns in video-recordings. Peak audio hour words were again between video and daylong audio ($M_{top\ audio}$=286.90(37.94); $M_{same\ audio}$=210.40(26.72)). 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.

---

[7]The same pattern held with video compared to peak audio hours.

Discussion

Our results can be distilled to three key findings. First, the density of noun input in hour-long video recordings was more similar to peak times in daylong audio recordings rather than to the day at large. Per minute, infants heard ~2–4x more noun input across quantity, speaker, utterance-type, and object-presence measures when video-recorded for an hour versus audio-recorded for a day. Second, while our metrics generally correlated across recording-types and many gross patterns were conserved across them, audio- and video-recordings differed in the relative rates of the top utterance types. That is, videos featured more questions and fewer declaratives than audio-recordings did. Finally, while the highest frequency words across recording types largely overlapped and correlated, top words from the daylong audio-recording appear to better represent the noun input across families.

Noun Quantity and Lexical Diversity

Overall, the pattern across recording-types primarily suggests a difference in volubility, since by-and-large, measures both correlated and differed quantitatively by recording-type. As Suskind et al.(2013) noted regarding interventions, daylong audio-recordings likely provide more realistic counterparts to "best behavior" hour-long videos. We add that shorter video-recording itself may influence volubility, resulting in samples more akin to the high points in the natural ebb and flow over the day.

Indeed, families likely found it simply easier to behave freely with infants in special vests than with cameras on their heads. Our finding that both "hat" and "camera" were top-10 video words supports this idea; no analogous nouns (e.g. vest, recorder) topped the audio-recording frequency rankings (see Figures 5 and 7). Anecdotally, while infants often required coaxing to wear the video-recording gear, no such issues emerged for audio-recording.

Given that we held family and age constant, we expected many similarities across recording-types; nevertheless, differences also emerged. Indeed, the quantity metrics provide

a conceptual replication and extension of Tamis-LeMonda et al. (2017). Despite numerous methodological variations (recording types and lengths, experimenter presence, age, word-class), both studies found that parent talk per unit time was significantly higher in shorter recordings on average, but lower than the *highest* portion of the longer recordings. This general pattern appears robust across our sampling methods. Taken together, this suggests that shorter recordings elicit denser, though not maximal, caregiver talk compared to what infants typically experience.

For certain research questions, such quantity differences may not matter, e.g. for studies examining *relative* word rates or object interactions in concentrated in-lab exposures. In contrast, research quantifying language input across populations with varying demographic, social, and cultural properties may need to be particularly sensitive to cross-sample comparison (cf. Bergelson et al., under review; Cristia, Dupoux, Gurven, & Stieglitz, 2017; Shneidman & Goldin-Meadow, 2012).

Object Presence

Rates of object presence were higher in videos than daylong audio-recordings, but equivalent between videos and peak audio times. Given that object presence correlated across recording-types within children (τ=0.40), our interpretation is that during higher talk volume times (i.e. video recordings and peak audio hours), nouns did occur with more object presence (i.e. infants mostly stayed in 1–2 rooms, interacting with what was at hand). However, since object presence was coded based on linguistic context and, when available, visual context, it's possible that indistinguishable object presence across video and peak audio is due to a combination of noise and systematic bias in coding object presence without visual context. Because object presence and the related ideas of referential transparency and contingent talk have been linked with early language development based on both audio-only and video-recordings (Bergelson & Aslin, 2017; Cartmill et al., 2013; McGillion, Pine, Herbert, & Matthews, 2017; Yurovsky et al., 2013), we find this latter possibility somewhat

unlikely. Indeed, a better understanding of what elicits contingent, referentially-transparent caretaker talk may be a fruitful avenue for further work.

Talker Variability

Infants heard nouns from more talkers per minute in videos than in daylong audio-recordings. In contrast, infants heard roughly double the speakers over the course of a day as they heard in one video-recorded hour and significantly more talkers during peak audio times than during videos (see Fig. 1).

Notably, while we considered noun input from all sources, 65.80% of infants' input came from mothers. Here, peak audio and video input from mothers was equivalent, though in comparison with daylong audio, there again were more nouns per minute from mothers in videos. In contrast, input from fathers was the only measure that did not vary in videos vs. daylong audio-recordings in the normalized count data. However, in the peak audio hours, there were more nouns from fathers than in the videos. Relatedly, >50% of videos captured *no* input from fathers. We believe this is because video-recording took place during weekday business hours when fathers in this sample were largely at work, while audio-recordings spanned work-hours and days. Given that fathers and mothers contribute differentially to early language development (Pancsofar & Vernon-Feagans, 2006), this is a clear example of a consequence of methodological choices. To better understand parents' input, considering work-schedules is critical. Put otherwise, home-recordings scheduled at the researcher and primary caretaker's convenience will likely undersample other caretakers.

The present results suggest that while infants hear most of their input from their mothers, they also hear several other speakers during high talk-volume times. Such data in turn feed infants' word-form representations. Indeed, recent lab studies have found that at the same age tested here, infants looked equivalently to named images when words were produced by a new person or their mother (Bergelson & Swingley, 2017), suggesting some degree of cross-talker normalization is in place around 6 months. In contrast, 14-month-olds'

learning of similar-sounding words improves after training with tokens from multiple speakers (Rost & McMurray, 2010), suggesting that even small amounts of talker variability aid new learning. This dovetails nicely with recent work showing that talker variability differentially influences certain phonetic discriminations (Bergmann et al., 2016). While a wide range of talker and token distributionss surely result in appropriately language-specific phonetic categories, we suggest that learning models incorporating a large dose of input from a single talker alongside smaller doses of input from 3+ other talkers may help inform word-form knowledge in infants similar to those tested here.

Utterance-Types

Per unit time, we found more nouns in every utterance-type in videos than in daylong audio-recordings. In particular, we did not anticipate differences in declaratives and questions. Indeed, while these utterance-types universally made up the majority of noun input, videos had relatively more questions and fewer declaratives. This is a key instance where methodological choices may influence language acquisition theories: base rates of questions taken from videos would inflate estimates of auxiliary verbs in the input. Notably, previous work has varied in whether links between questions in the input and children's early productions emerged, with developmental level invoked to explain cross-study differences (Barnes, Gutfreund, Satterly, & Wells, 1983; cf. Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002). We add the possibility that recording-type too may contribute to the base rates of questions in the input, even with age and recording length kept constant.

Top Words

Our third key finding concerned noun frequency and commonality across families. We found that top words in the daylong audio-recordings were heard by $\geq 84\%$ of families; only 1/10 top video words ("hat") was this common, a clear vestige of our recording equipment (see Figure 7). This result may be meaningful in several ways. First, our analysis suggests that the input would seem far more heterogeneous across children based on hour-long

video-recordings than it really is. Second, word frequency and prevalence are often used to select stimuli for in-lab study; relying on estimates from shorter, less representative recordings may stymie the words studied in the lab. Thus, understanding how cross-family noun-input stability scales with recording-length and type 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 on battery life for small video-recorders, we cannot conclusively separate the effects of modality and length. That is, had we recorded daylong videos, we may have obtained equivalent results across recording-types. Indeed, our peak audio analyses provided some evidence that videos are more akin to particularly language-saturated parts of infants' experience. However, the peak audio comparisons are imperfect since these hours were not bookended by researchers arriving and departing with pesky gear; further comparisons await technological progress. Importantly, we do not mean to suggest that audio reigns supreme: for many language-relevant questions concerning gaze, gesture, and visual perception, it is simply insufficient.

A further limitation here is self-selection: many parents are unwilling to invite home recordings. Relatedly, our participants do not reflect U.S. demographics (let alone those elsewhere), and should be extended 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. These are first steps in unpacking how two different data collection approaches may influence conclusions about early linguistic input, with a narrow focus on the initially dominant lexical class of nouns. 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 these factors matter for early learning. Yet without knowing how sampling methods may constrain results, we necessarily limit adequate models of language acquisition. The present work charts datapoints within this largely underspecified space, probing the robustness of linguistically-relevant measures across naturalistic sampling methods of infants' everyday experiences.

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. (2017). Child-directed speech is
infrequent in a forager-farmer population: A time allocation study. *Child
Development*.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior
Research Methods, Instruments, & Computers*, *28*(1), 125–127.

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.

Mindell, J. A., Sadeh, A., Wiegand, B., How, T. H., & Goh, D. Y. (2010). Cross-cultural differences in infant and toddler sleep. *Sleep Medicine, 11*(3), 274–280.

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.

Table 1

Infant ages at home recordings and enrollment lab visit

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

Table 2

Count measures (*n*=12), by Measure-Type

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

Table 3

Proportion of infants with no recorded nouns for the listed measures, by sample

Time Sample	Fathers	Mothers	Reading	Singing	Imperatives
A	NA	NA	0.16	0.02	NA
SameA	0.27	NA	0.43	0.11	NA
TopA	0.18	0.02	0.27	0.07	NA
V	0.51	0.09	0.34	0.11	0.02

Note. V=video, A=daylong audio, Top=top hour of A, Same = Video-hour of A. All infants heard nouns for all other measures (see Table 2).

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	0.44 (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
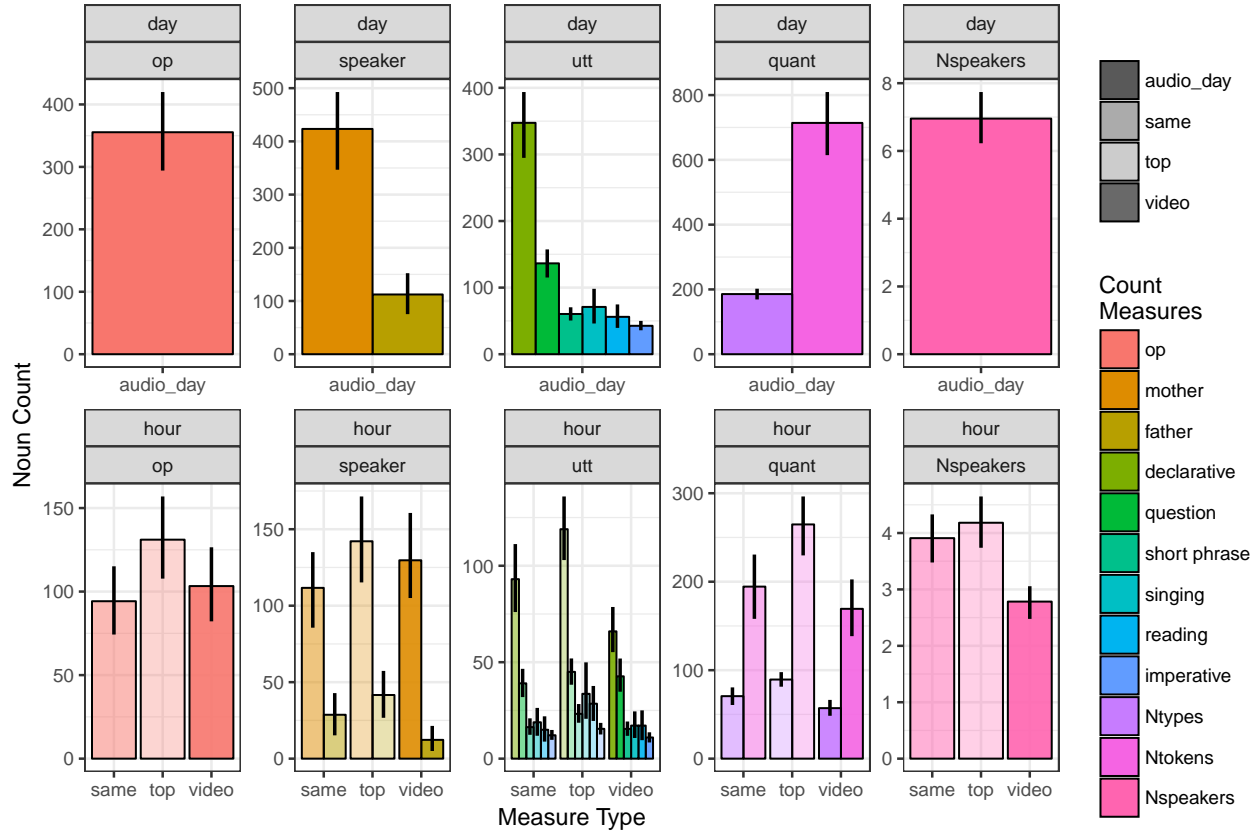
Figure 1. Noun count measures across audio-recordings and videos. Top row depicts daylong audio data; bottom row shows the 3 hour-long annotations: "same" and "top" are the two peak audio times, and "video" indicates the video data. Upper panel labels indicate annotated sample length (day or hour); the bottom panel labels reflects measure type (op = object presence; utt = utterance-type, quant = quantity, Nspeakers = number of speakers). Bars (left to right) appear in legend order (top to bottom) in both color (count measures) and opacity (time sample: day, top-hour, same-hour, or video).
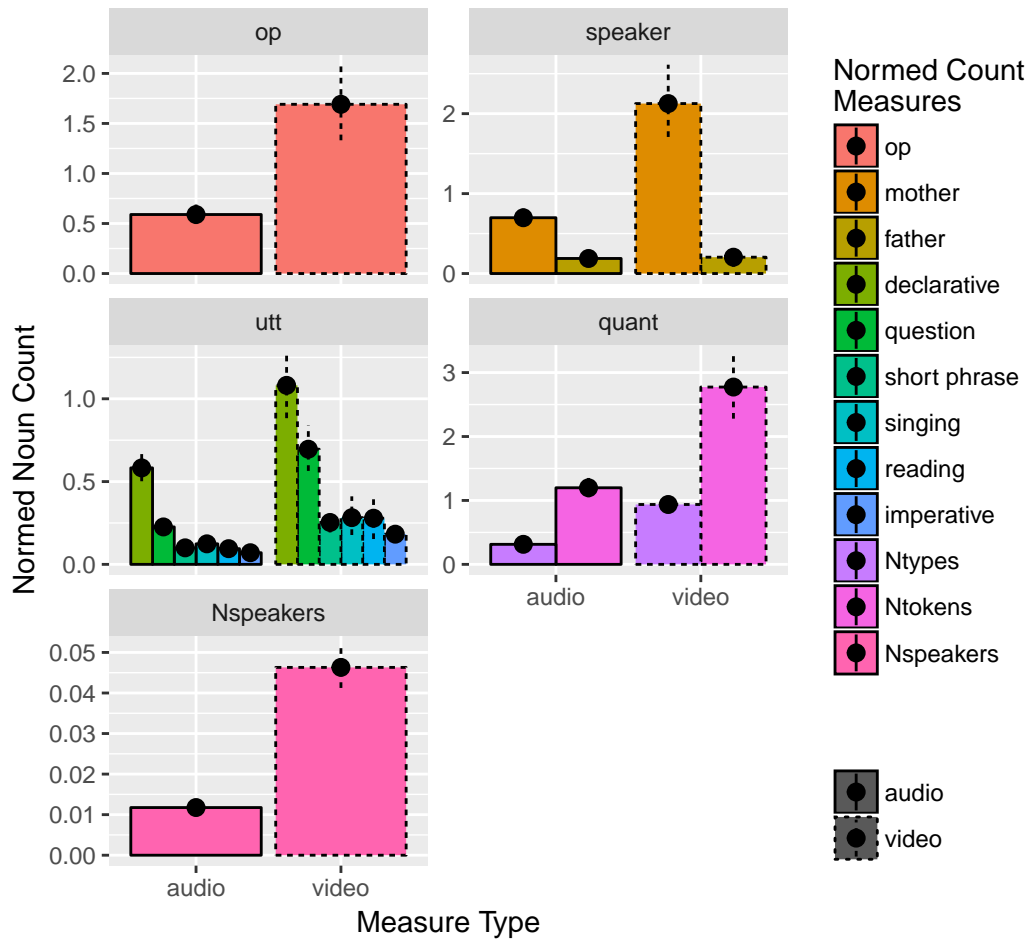
Figure 2. Noun count measures normalized by recording length, for audio-recordings (solid borders) and videos (dashed borders). Normalized counts were calculated by dividing raw counts (see Fig 1.) by non-silent recording minutes. op = object presence; utt = utterance-type, quant = quantity, Nspeakers = number of speakers. Bars (left to right) appear in legend order (top to bottom). All measures differed significantly across recording-types except nouns from fathers.
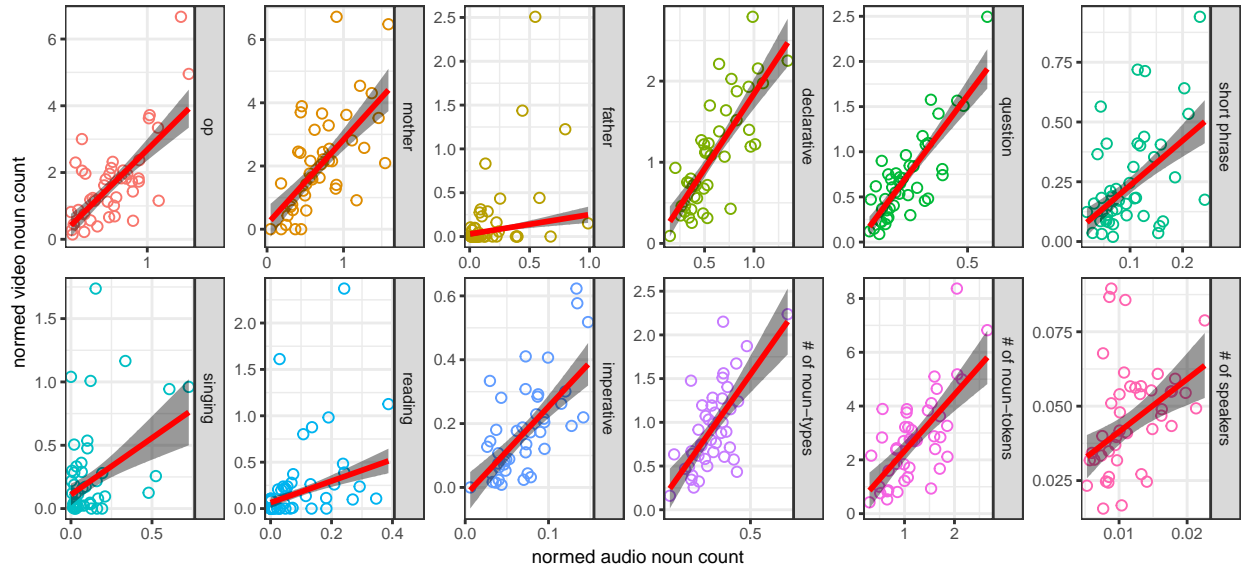
Figure 3. Normalized count correlations between audio- vs. video-recordings. Each point indicates nouns per minute of recording for each child, averaged across months 6 and 7, for each measure. Point-shape indicates measure type. Robust linear correlations are plotted for visualization only; non-parametric correlations (Kendall) were computed for analysis, showing that all correlations were significant except nouns from fathers and in singing.
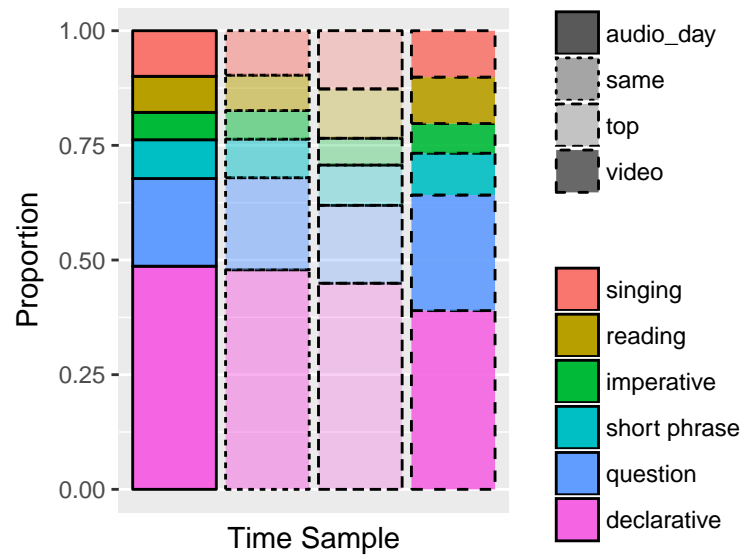
Figure 4. Utterance-type proportions across audio-recordings (daylong, "same" hour and "top" hour) and videos (indicated by line-type). Utterance-types are in legend order top to bottom. Videos contained a significantly more questions and fewer declaratives than the audio-recording time samples.
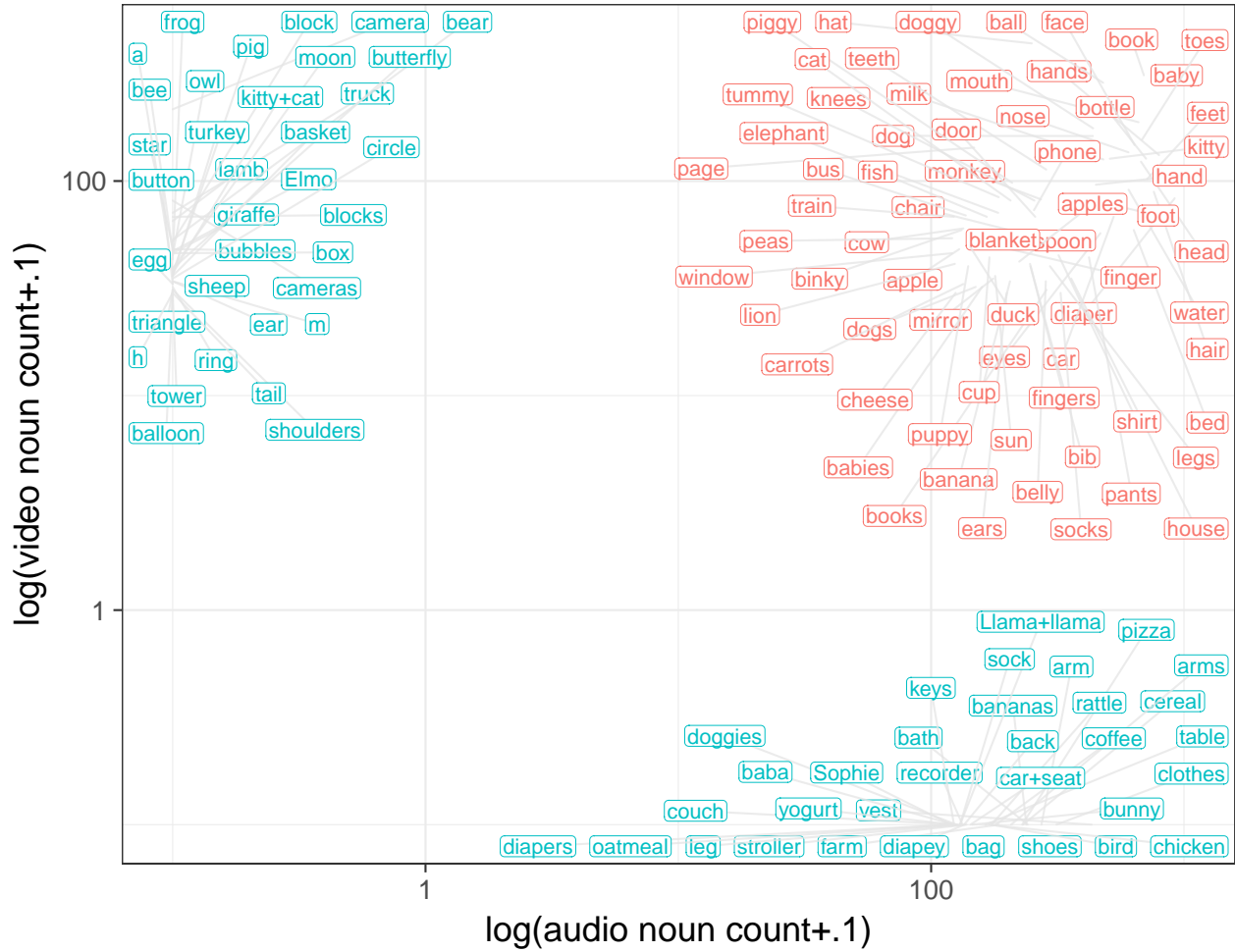
Figure 5. Log-scaled counts of the top 100 words in audio- and video-recordings. Each node represents the averaged count, across all participants in both months, of each noun (0.1 was added before taking logs to include 0 counts.) Words in blue occurred 0 times in one recording type; words in pink were attested in both recording types. Nodes are jittered for visual clarity, with grey lines indicating node location on axes.
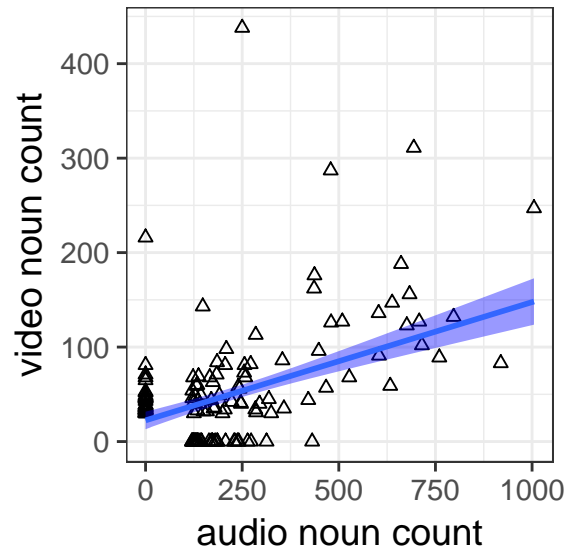
Figure 6. Correlations of the frequencies of the top 100 words in audio- vs. video-recordings.

Each node represents one word averaged across all participants in both months.
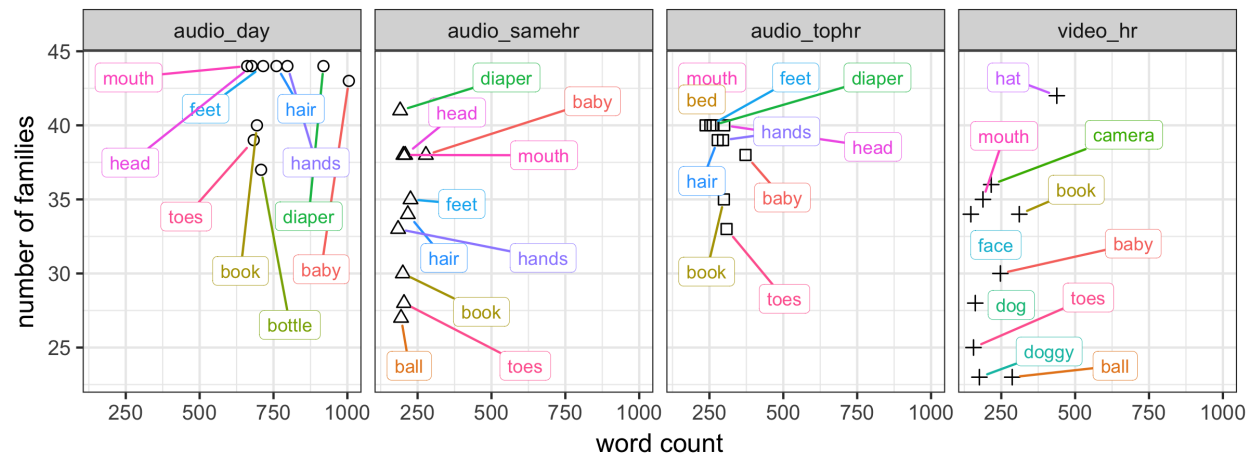
Figure 7. Top 10 words by recording type and time sample. Each node represents the frequency count of each top audio or video word over both months (x-axis) and the number of families where that word was said (out of 44) across months (y-axis).