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[2], & Shaelise Tor[2]

[1] Duke University

[2] University of Rochester

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Elika Bergelson, Postal address. E-mail: elika.bergelson@duke.edu

Abstract

Measurements of infants' experiences in their typical daily environments provide critical information about early development. However, the role of sampling methods in providing this information is rarely examined. Here we directly compare language input measures collected by hourlong video- and daylong audio-recording within the same group of 44 infants, at 6 and at 7 months. Month to month, our results are incredibly consistent, suggesting our methods obtain reasonable estimates of nouns in the input across the month-long gap. However, comparing recording-types, we find large differences across language quantity and lexical diversity, talker variability, utterance type, and referential transparency. Put briefly, video recordings featured far denser quantities of nouns, relatively more input from mothers, more questions and declaratives, and more talk about present objects. This suggests short video-recordings may inflate language input estimates, and should be used cautiously for extrapolation about common words, talkers, and situational/contextual features at larger timescales. Analyses based on counts and proportions often led to different conclusions; we suggest proportions may provide ill-advised compression of the information available to infants. Even when measures are standardized per unit time, or computed proportionally, hourlong video and daylong audio-recordings provide a fairly divergent picture of the input infants hear, and therefore learn from, in their daily lives. If our theories are to be held accountable to our observations, we suggest greater care be taken to unpack the ramifications of underlying methodological and analytic decisions for measuring infant language input.

Keywords: keywords

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

Introduction

Researchers have studied development by observing infants experiencing their natural habitats for over a century (Taine, 1876) (**ADD REF**: Williams, 1936). 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, B. C. Roy, Frank, DeCamp, Miller, and Roy (2015), Oller et al. (2010), Weisleder and Fernald (2013), VanDam et al. (2016)) (**CHECK REF**: Bergelson & Aslin, 2017b,, Roy et al., 2015, Oller et al., 2010, Weisleder & Fernald, 2013, VanDam et al, 2016, inter alia.) 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, 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) (**CHECK REF**: Tamis-Lemonda et al.

(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, in structured interactions infants generally experienced more language both in word-quantity (i.e. tokens) and word-variability (i.e. types) per minute. They also found that language quantity across contexts correlated, and that the peak five-minutes of the naturalistic interaction was 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 cautiously.

In contrast, work by Hart and Risley (1995) (**CHECK REF**: 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 (Fernald, Marchman, & Weisleder, 2013) (**CHECK REF**: Noble et al, Fernald et al, 2013, inter alia), they have been criticized as an extreme over-extrapolation (Dudley-Marling & Lucas, 2009) (**CHECK REF**: Dudley-Marling and Lucas, 2009; **ADD REF**: 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) (**CHECK REF**: Tomasello, 2000; Lidz et al, 2003; Brent and Siskind, 2001). Here, the research question dictated what was 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"). But for questions concerning infants' language input, it is largely unknown how methodological choices may

bias our answers.

In the present study, we explore these issues, directly comparing hour-long video-recordings and daylong audio-recordings in a single sample of 44 infants, at 6 and at 7 months, 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 (Hoff & Naigles, 2002, Brent and Siskind (2001), DeBaryshe (1993)) (2) referential transparency, which clarifies whether the target of the spoken word is visually appreciable (Bergelson & Aslin, 2017,Bergelson and Swingley (2013), Yurovsky, Smith, and Yu (2013)) (**ADD REF:** Trueswell et al), and (3) talker, which lets us quantify the range of speakers infants hear (Rost & McMurray, 2010, Bergmann, Cristia, and Dupoux (2016)) (**CHECK REF:** Rost & McMurray, Bergmann et al, 2016, Cogsci).

This design sets up two overarching questions. First, we examine extrapolative validity: how well do the data from one video-recorded hour predict the absolute quantity and relative distribution of data in an entire audio-recorded day? Separating quantity and distribution is important given that (a) one may scale more robustly with recording length than the other and (b) we simply do not know how infants themselves aggregate their input. That is, count-based and proportional metrics may prove differentially predictive of language learning. Indeed, certain linguistic metrics like type-token ratio (a lexical-diversity metric) scale poorly with sampling length (Covington & McFall, 2010), but yet, along with their "absolute" counterparts (type and token counts), predict language development (e.g. Hoff & Naigles, 2002, Pan, Rowe, Singer, and Snow (2005),Rowe (2012),Hoff and Naigles (2002)) (**ADD REF:** Huttenlocher et al, 2001;2010; **CHECK REF:**Hoff and Naigles, 2002). For others metrics, we simply do not know how they scale with recording-length, or whether relative or absolute quantities predict learning better (or even differentially). Here we chart some points within this underspecified space, probing how robust linguistically-relevant measures are

across two sampling methods of infants' everyday experiences.

Second, we assess input-stability within sampling method: do infants receive quantitatively different language input when we audio- or video-record at 6 months compared to when we do so four weeks later? Put otherwise, are there effects of "initial" vs. "subsequent" observational recordings, and do these vary by recording length and modality? Given that there are no major developmental milestones between 6 and 7 months, we predict strong convergence across time-points. Several accounts are compatible with cross-month differences, including the possibility that caregivers simply behave differently at initial versus subsequent home visits.

Thus, our main goal was to compare language input young infants receive across four key properties (frequency, utterance-type, referential transparency, and talker), as measured by an hour of video and a (separate) full-day audio-recording, each at two time-points. This seemingly methodological question has deep implications for developmental theory: we examine how sampling and aggregation approaches may alter 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 $\pm$ 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; see table XX.[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 analysis, are reported elsewhere (Bergelson & Aslin, 2017b; Laing and Bergelson, under review). Four recordings were collected for each infant: an audio- and video-recording at six and at seven months. Each recording was on a different day. See table XX.

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 life 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 by Bluetooth with an iPad/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.

[1] We include only these timepoints because at this stage of development no infants had begun producing words themselves (which may change the input for reasons orthogonal to those we examine here), and given the broader project aims, these timepoints alone had the entire daylong audiorecording annotated.

Audio-Recordings

Audio-recordings captured a full day (up to 16 hours) of infants' language input. Parents were given vests with a small chest-pocket, and LENAs (LENA Foundation, Boulder, CO), small audio-recorders (<60g) that fit into the vest pocket. 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. 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. 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 next annotated by trained researchers. The "sparse annotation" entailed marking each concrete noun heard by the child. This includes 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 lemmatized to it's "basic level" or 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.

We assessed intercoder reliability on a random contiguous 10% of the annotations in each file.

Results

Analysis Plan

Based on the coding scheme above, we derived count (n=12) and proportional (n=10) measures from each recordings' annotations for each child (n=44), recording-type (audio, video), and month (six, seven). See Table XX. We also normalized the count measures by recording length; further details are below. We initially created multi-level models, with each of the 22 measures as the dependent variable, recording-type and month as fixed effects, and participant as a random effect (i.e. dv ~ recording_type + month + (1|subj)). Month was not a significant predictor in any of the 22 models; recording-type was in 17/22 models (see S.I.). However, since many of the models showed structured (generally funnel-shaped) residuals that limited interpretation across measures, we instead report a simple set of nonparametric analyses.

For all recording type and month 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 each set of Wilcoxon

and Kendall tests. For instance, comparing the 12 count measures for month six vs. seven within audio recordings by Wilxocon test is one "set". With our analysis plan of 2(count or proportional) x 2(audio or video) x 2 (six or seven month) x 2(Wilcoxon Test or Kendall correlation), we applied this p-value adjustment to 15 further sets.

Count Measures, Month 6 vs. 7

We first analyzed the 12 count measures in month six versus seven, by recording-type. Across children, none of the 12 differed by month within audio recordings (all adjusted-p>.05), or within video recordings (all adjusted-p>.05.) Testing the correlations, all 12 count measures correlated significantly month-to-month for audio (Kendall's tau ranged from 0.29 - 0.52, all adjusted-p<.05), and 11/12 did so for video (Kendall's taus ranged from 0.29 - 0.60, all adjusted-p<.05), excluding number of nouns heard in reading (adjusted-p>.05). Thus, within recording type, the count-based metrics of the object words infants heard were statistically equivalent in month six and seven, and correlated significantly within children month-to-month (except 1/24 correlations); see Figure 2. This suggests that parents are acting naturally, or at least consistently, during our home recordings each month.

Count Measures, Audio- vs. Video-recordings

We next assessed our count measures across recording-types, examing how noun input scaled between hour-long video-recordings and daylong audio-recordings. Modally, videos were an hour (62 min, M=60.79, SD=6.31, R=27.9-74.9 min), and audio-recordings were 16 hours (960 min, M=858.41, SD=119.41, R=635-960 min), the maximum capacity of the LENA device. While determining the exact onsets and offsets of naps from audio alone is not possible, by removing the "silent" portions of the recordings (see Methods), we estimated an upper-limit on infants' awake (i.e. non-silent) time (Mode = 10.90, M = 603, SD=106.8, R=385.2-951 min). This comports with established norms for 6–8-month-olds in the US (**ADD REF:** Mandel et al, 2010), which are 3 hours of daytime sleep, and 10 hours of nighttime sleep. Thus, 16 hours of recording beginning when the child wakes up should

capture ~11 daytime waking hours, in line with our silence demarcations. 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).

```
## # A tibble: 2 x 16
##    month aboost_min aboost_awakemin aboost_types aboost_tokens
##   <fctr>      <dbl>           <dbl>        <dbl>         <dbl>
## 1     06      13.89            9.88         4.05          5.79
## 2     07      14.82           10.45         4.65          6.03
## # ... with 11 more variables: aboost_speakers <dbl>, aboost_MOT <dbl>,
## #   aboost_FAT <dbl>, aboost_d <dbl>, aboost_q <dbl>, aboost_i <dbl>,
## #   aboost_s <dbl>, aboost_r <dbl>, aboost_n <dbl>, aboost_op <dbl>,
## #   comp <chr>

## # A tibble: 2 x 16
##    month aboost_min aboost_awakemin aboost_types aboost_tokens
##   <fctr>      <dbl>           <dbl>        <dbl>         <dbl>
## 1     06       2.19            1.75         2.46          4.06
## 2     07       4.19            3.99         3.63          5.10
## # ... with 11 more variables: aboost_speakers <dbl>, aboost_MOT <dbl>,
## #   aboost_FAT <dbl>, aboost_d <dbl>, aboost_q <dbl>, aboost_i <dbl>,
## #   aboost_s <dbl>, aboost_r <dbl>, aboost_n <dbl>, aboost_op <dbl>,
## #   comp <chr>

## # A tibble: 2 x 16
##    month vboost_min vboost_awakemin vboost_types vboost_tokens
##   <fctr>      <dbl>           <dbl>        <dbl>         <dbl>
## 1     06       0.07             0.1         0.35          0.31
## 2     07       0.07             0.1         0.30          0.25
```

```
## # ... with 11 more variables: vboost_speakers <dbl>, vboost_MOT <dbl>,
## #   vboost_FAT <dbl>, vboost_d <dbl>, vboost_q <dbl>, vboost_i <dbl>,
## #   vboost_s <dbl>, vboost_r <dbl>, vboost_n <dbl>, vboost_op <dbl>,
## #   comp <chr>

## # A tibble: 2 x 16
##    month vboost_min vboost_awakemin vboost_types vboost_tokens
##   <fctr>      <dbl>           <dbl>        <dbl>         <dbl>
## 1     06       0.01            0.02         0.23          0.31
## 2     07       0.01            0.02         0.14          0.15
## # ... with 11 more variables: vboost_speakers <dbl>, vboost_MOT <dbl>,
## #   vboost_FAT <dbl>, vboost_d <dbl>, vboost_q <dbl>, vboost_i <dbl>,
## #   vboost_s <dbl>, vboost_r <dbl>, vboost_n <dbl>, vboost_op <dbl>,
## #   comp <chr>
```

To examine how the hour-long video data "scale" to day-length data descriptively, we divided the 12 audio count metrics by the 12 video count metrics for each child, to derive "audio-boost" scores. This showed that the audio-recordings were ~14x longer than the videos, or 10x longer if only "non-silent" portions of the audio-recording are included. However, rather than a concomitant 10-fold increase in our count metrics (as would be expected if the video captured a "representative" hour of the day), the boost was closer to fivefold across measures; see Table XX. This suggests that the videos, by and large, had a denser concentration of nouns across our measures than did the audio recordings.

Notably, "zero" values (e.g. recordings in which there were no nouns heard in singing) were omitted from the audio-boost computations given that they result in undefined values for a given child in a given month. The majority of variables had at least one such value, with over 1/3 of video recordings lacking instances of nouns heard in singing, reading, or by fathers each month; see Table 2.

Table 1

month	v_MOT	a_FAT	v_FAT	v_q	v_i	a_s	v_s	a_r	v_r	v_n
06	0.16	0.05	0.52	0.00	0.07	0.07	0.32	0.25	0.55	0.02
07	0.09	0.09	0.75	0.02	0.07	0.07	0.34	0.25	0.50	0.07

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. Unlike for the audio-boost calculation, this allows us to retain zero values, rendering more readily interpretable results across our count and proportional measures.

With the normed data, we found identical patterns when looking within month six and seven: /12 of our count-based metrics (in each month) occurred at significantly lower rates in audio recordings than video recordings (adjusted-p<.05). The three that were statistically indistinguishable were nouns/minute produced (1) by fathers, (2) in reading, and (3) in singing (all adjusted-p>.05). These same three measures had a large number of zero values; see table XX. We return to the topic of unattested low-frequency events in corpora in the discussion.

The pattern of correlations across recording-types was mixed: in month six, audio vs. video normed count data only correlated significanly for 3/12 metrics, which were all utterance-type measures: declaratives (tau = 0.44), questions (tau = 0.33), and imperatives (tau = 0.30; all 3 adjusted-p<.05). For month seven, 11/12 metrics correlated in audio vs. video data; number of nouns per minute heard in singing did not (excluding singing, 0.27 - 0.53, all adjusted-p<.05). See Table 3 and Figures 3 and 4.

Table 2

v_MOT	a_FAT	v_FAT	v_i	a_s	v_s	a_r	v_r
0.09	0.02	0.52	0.02	0.02	0.11	0.16	0.34

Table 3

norm_meas	06	07
y_op	2.90	2.80
MOT	2.90	3.20
FAT	1.50	0.80
d	1.90	1.80
q	3.10	3.00
n	2.90	2.20
s	1.80	3.00
r	4.00	2.10
i	2.40	2.80
numtypes	3.10	2.90
numtokens	2.40	2.30
numspeakers	4.40	3.60

Proportion Measures, Month 6 vs. 7.

Turning to the 10 proportion measures, as with the count measures, there were no significant differences between month six and seven within audio- or within video-recordings (all adjusted-p>.05). The pattern of correlations differed somewhat from the count measures: for the audio-recordings, 7/10 proportional measures correlated from month 6 to 7 (for these seven: Kendall's tau range: 0.31-0.46, adjusted-p<.05; for object co-presence, questions, and short-phrases adjusted-p>.05). For videos, only proportion of input from mom and from dad correlated significantly from month 6 to 7 (Kendall's tau = 0.45 and 0.63, respectively; adjusted-p<.05). Thus, overall, the proportional metrics were indistinguishable month-to-month within recording-type, but the correlations between the proportional measures across children at month six and seven were variable, especially for video-recordings.

Proportion Measures, Audio- vs. Video-recordings

Across recording-types, proportional measures differed more substantively than count measures; see figure XX. At six months, 4/10 proportional measures differed significantly between audio- and video-recordings (type-token ratio, and the proportion of object presence, declaratives, and questions, all adj-p<.05), while at seven months, 6/10 did so (the same four as in month six, plus input from mothers and fathers; all adj-p<.05). See Figure XX.

Descriptively, videos featured object presence more than audio-recordings (M_{Δ} = %), likely reflecting a narrower range of activities during video recording. Videos also had a higher type-to-token ratios (i.e. more lexical diversity): M= vs. for audio-recordings, consistant with previous work (Covington & McFall, 2010).

Across utterance-types, the overall distributions of nouns in audio- and video-recordings were similar: the majority of the noun input was declaratives, followed by questions, with the remaining input spread across imperatives, reading, singing, and short-phrases; see Figure XX. However, audio-recordings featured relatively more

declaratives and fewer questions than videos (, and %, respectively). Finally, in month seven, mothers talked % less in audio than video, and fathers talked % more.

Turning to correlations across recording-types, in month six only the proportion of imperatives reached significance (Kendall's tau =, adj-p<.05). At seven months, three of the variables that differed significantly also correlated significantly (proportion of noun input from fathers and mothers, and in declaratives), along with the proportion of nouns heard in reading, (all four adj-p<.05, Kendall's tau range: 0.30 - 0.35).

Noun Frequency and Prevalence

We conclude with a set of highly exploratory analyses at the word level, which aim to provide a first-pass characterization of whether audio and video recordings captured the same nouns and the same relative frequencies by examining word frequency across each month, recording type, and infant. The distribution of nouns in our recordings was zipfian: of the 5801unique object words (3137 lemmas) heard across months and recording types, only 2482 (960 lemmas) were heard more than once (see Figures 8 and 9).

Collapsing month, 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 10). Further, we found numerically stronger correlations month-to-month than across recording-types within month (month 6 audio vs. video: tau = 0.15, month 7 audio vs. video: tau = 0.15, month 6 vs. 7 audio: tau = 0.50, month 6 vs. 7 video: tau = 0.26, all adjusted-p<.05; see Figure 11). Thus, at the word-level too, month-to-month measures appear more stable than cross-recording-type measures

Finally, looking at just the top ten words by month and recording-type, we again find greater consistency across months than recording-types (see Figure XX and Table XX). Indeed, top words within recording-type were largely overlapping, while only two words

overlapped on all four lists (book and baby). Furthermore, characterizing how words were spread across infants and months, we find that the top audio words were far more common across families than the top video words were (see Figure 12). Thus, daylong audio-recordings appear to render more stable high-frequency words across families and across months than do video-recordings.

Discussion

Our results can be distilled to three key findings. First, the input was relatively stable month to month. That is, within recording-type, infants heard the same amount of concrete noun input across all of our metrics in month six and seven, using both count and proportion-based measures. Second, there was low extrapolative validity across recording-types. Per minute, infants heard ~2-5x 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 an entire day. Proportional versions of the measures (which standardize over quantity of noun input rather than recording length) revealed smaller differences, in roughly half the measures. Finally, while the highest frequency words across recording types and months correlated (and exhibited Zipfian frequency distributions), top words from the daylong audio-recording appear to better represent the noun input across families.

Input Stability

Given our initial question concerning input-stability over the month-long interlude between recordings, we resoundingly find that within video-recordings, and within audio-recordings, there is little evidence for differences in noun input at six and seven months.

However, the correlations presented a somewhat different pattern across count and proportion measures. Almost all (23/24) correlations remained significant after p-value adjustment for the count-measures. Yet, for the proportional measures, within videos only

the two speaker-based measures (proportion of mother and father in the noun input) correlated across time; the other 8 measures did not. Even relaxing the p-value to the uncorrected .05 level, only two further measures correlated across months within videos: declaratives and imperatives. Within audio, while most proportional measures did correlate across time-points (7/10), the three that did not (object presence, questions, short-phrases) did correlate when calculated as counts instead of proportions. Given that these measure reflect variables whose prevalence is argued to influence language development (Bergelson & Aslin, 2017, Bergelson and Swingley (2013), Yurovsky et al. (2013), Brent and Siskind (2001), Swingley and Humphrey (2017), Hoff-Ginsberg (1985),Newport, Gleitman, and Gleitman (1977)) (**CHECK REF:** Yurovsky et al, Brent & Siskind; Newport et al, 1977), these results present a clear case where researchers would be led to different conclusions about cross-month stability based on whether count- or proportion-based analyses were used.

Extrapolative Validity

Our query into extrapolative validity (i.e. the comparison across recording-types) highlighted many differences across measures, even with family and age held constant. Across both time-normalized count and proportional metrics, measures of noun quantity, object presence, and the two most frequent utterance-type (questions and declaratives) differed significantly across recording-types. Indeed, by and large, only measures that occurred relatively infrequently across recordings did not vary in audio vs. video files (e.g. noun input in reading and singing, and from fathers).

The quantity differences we find (with video-recordings capturing more types and tokens per unit time, but also a higher type-token ratio) are consistent with previous research. Indeed, metrics of lexical diversity generally do not scale with corpus length, across various kinds of language samples (Covington & Mcfall, 2010). Our quantity results also conceptually replicate and extend those of Tamis-Lemonda et al (2017). Despite numerous methodological differences (recording lengths, experimenter presence, infant age, word class

analyzed), both studies find that parent talk per unit time is significantly higher in shorter recordings. While the difference they find is less extreme numerical (roughly 1.5-2x 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. To us this suggests that shorter recordings, in general, elicit denser caregiver talk.

We find consistently more object co-presence in video- 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, or additionally, it may be the case that there are more ambiguous cases of "object co-presence" in audio recordings than video recordings, which were in turn annotated as "not present" at higher rates. Given the XX rates of agreement, we find it more likely that this reflects a true difference between situations that arise during daylong-audio vs. hourlong-video recordings. Insofar as object presence is linked with early word learning (Bergelson & Aslin, 2017), a more extensive understanding of what modulates it is an important issue left to future work.

We did not anticipate that the top utterance-types would vary by recording-type. That is, 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 another example of a methodological choice 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,Furrow, Nelson, and Benedict (1979),Hoff-Ginsberg (1985)) (**CHECK REF:**see discussion in Huttenlocher et al, 2002). Here we add the possibility that recording-type too may contribute to the base-rates of questions in the input, even with child-age kept relatively constant.

Turning to the pattern of correlations across recording-types, as with time-point,

results were difficult to interpret. There was no measure that correlated between audio and video recordings consistently for both proportional and count-based measures, in both month six and seven. One possibility is that while rank-based correlations with p-value adjustment are an appropriate approach to protect against false positives, they likely also reduced our power to detect small correlations that may be present with this sample size. Alternately, it's possible that there simply aren't clear correlations in noun input across recording-types.

Top Words

Thus far, the overarching pattern in these results (stability in noun quantity month-to-month, but differences across audio versus video) suggests that parents behaved naturally, but that "natural" behavior differed by recording context. This is consistent with a point made by 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 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. It seems that families 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 that the equipment itself was more prominent in video-recordings, 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 ranking within or across families in our audio recordings (see Figure XX).

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' quotidien experiences. This underscores our third main result: that the conclusions one would draw about which words are most common in young infants' language input differed 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 was this common across families. This is true even though the video words had greater quantities of nouns per unit time; the top audio words only occured 2-4 times more often than the top video words (despite a 10-fold increase in awake recording time).

Limitations

Given the technical limitation that available infant-friendly video-recorders have a shorter battery life than audio-recorders at present, 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 (**ADD REF:** cf Bergelson et al, under review).

Conclusions

Understanding what infants learn from is a key part in 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.

Without knowing how our sampling methods may be limiting us in principle, we necessarily limit our ability to adequately model infant learning. We...

Notes to self: Might want to reorder points, and make point 3 a little more general, some overarching version of the differences in conclusions Discuss zeros/unattested datapoints

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.

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.

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.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, *17*(2), 94–100.

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.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language*, *6*(3), 423–442.

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.

Hoff-Ginsberg, E. (1985). Some contributions of mothers' speech to their children's syntactic growth. *Journal of Child Language*, *12*(2), 367–385.

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

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, id rather do it myself: Some effects and non-effects of maternal speech style.

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.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*(4), 763–782.

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.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*(5), 1762–1774.

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.

Swingley, D., & Humphrey, C. (2017). Quantitative linguistic predictors of infants' learning of specific english words. *Child Development*.

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*.

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.

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.

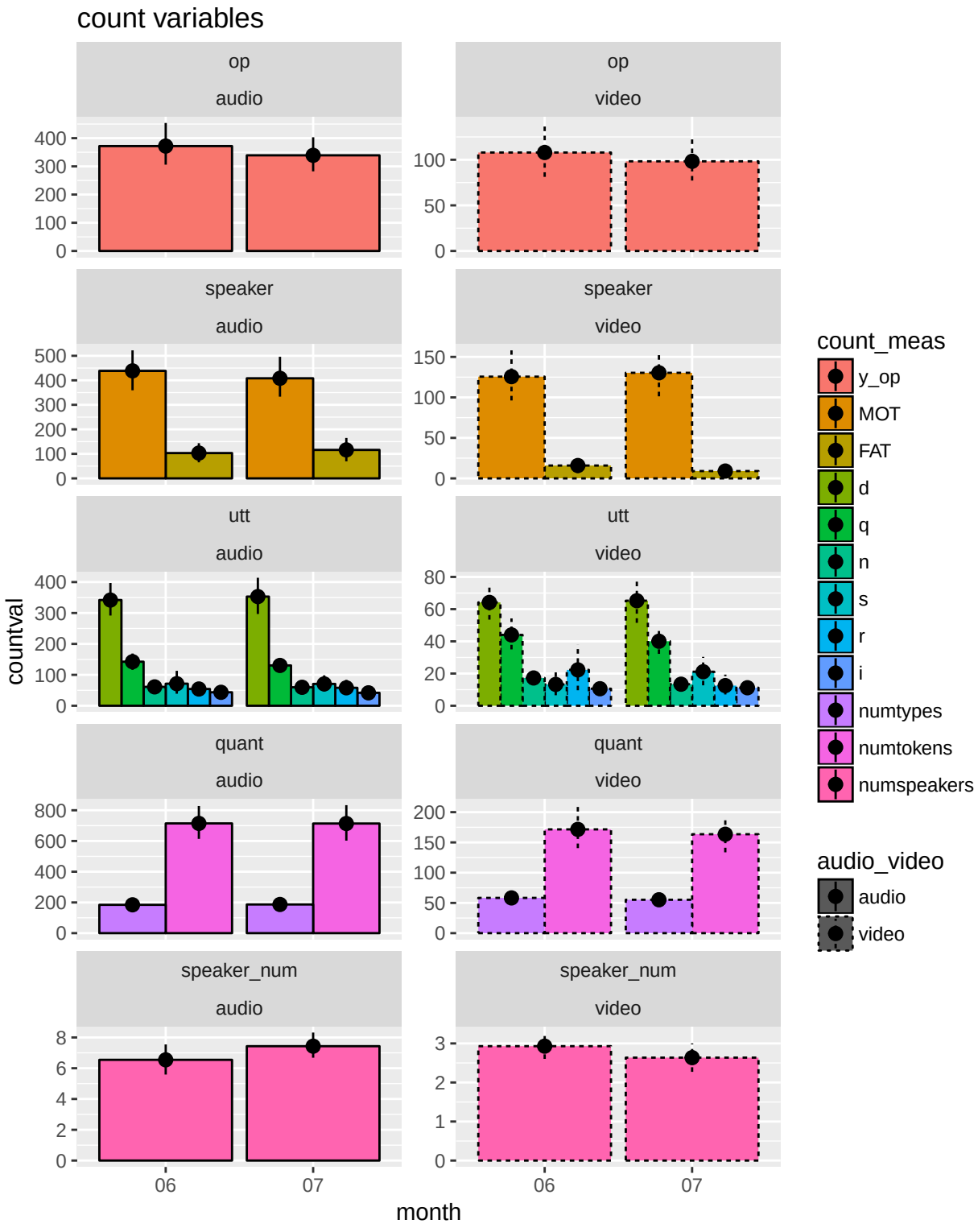
count variables
op
audio
op
video
speaker
audio
speaker
video
utt
audio
utt
video
quant
audio
quant
video
speaker_num
audio
speaker_num
video
countval
month
06
07
count_meas
y_op
MOT
FAT
d
q
n
s
r
i
numtypes
numtokens
numspeakers
audio_video
audio
video

Figure 1

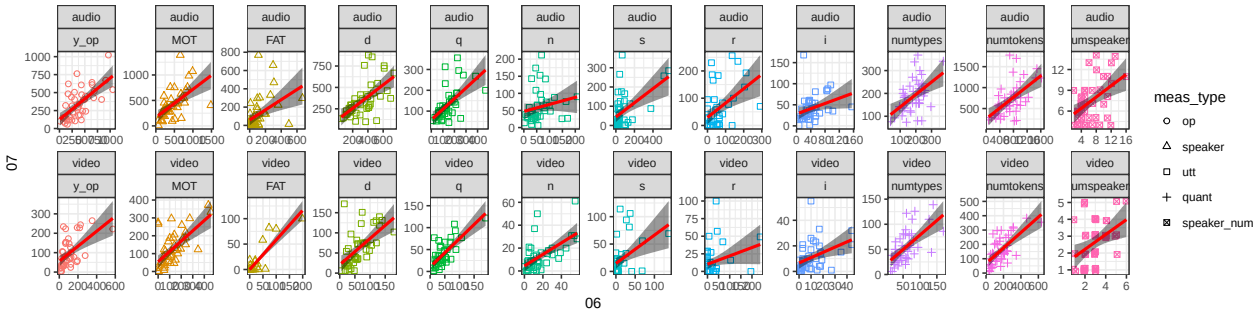

Figure 2. Count correlations

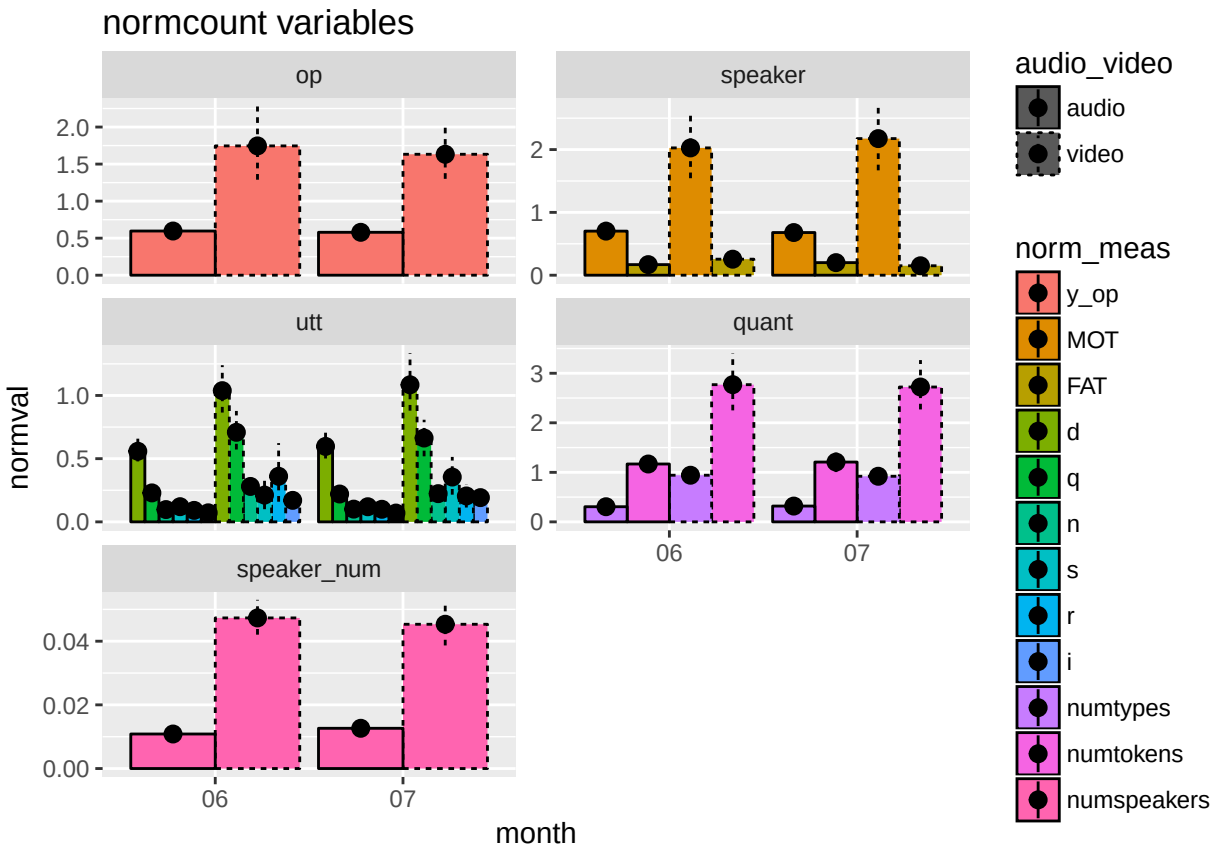

Figure 3. Normalized variable counts by month

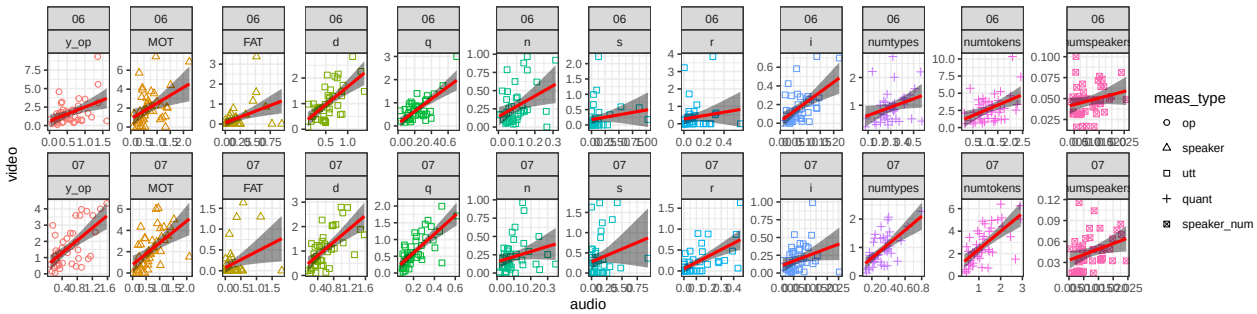

Figure 4. Normalized count correlations

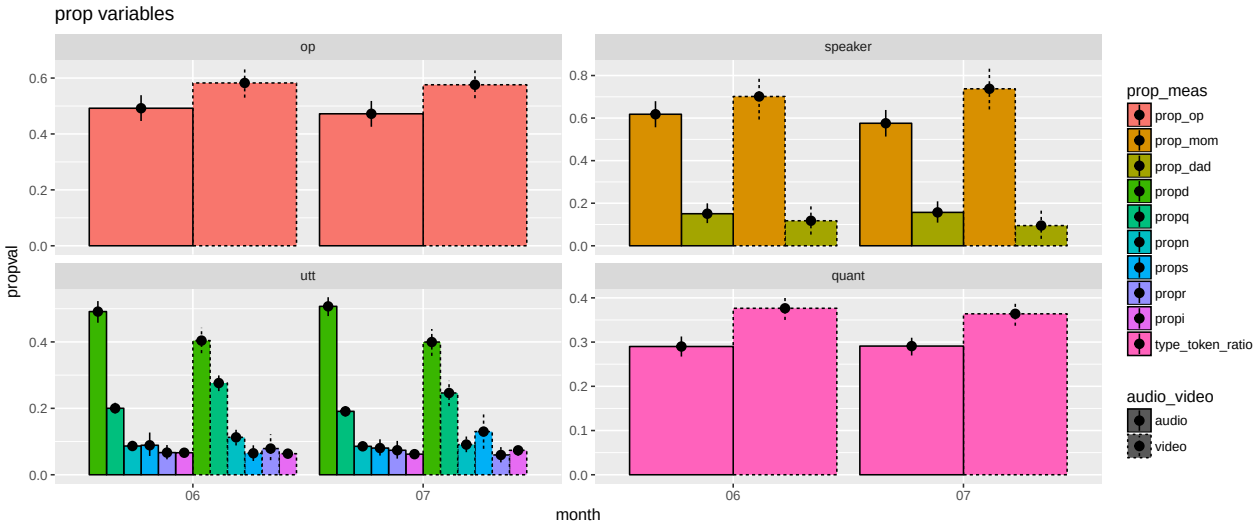

Figure 5

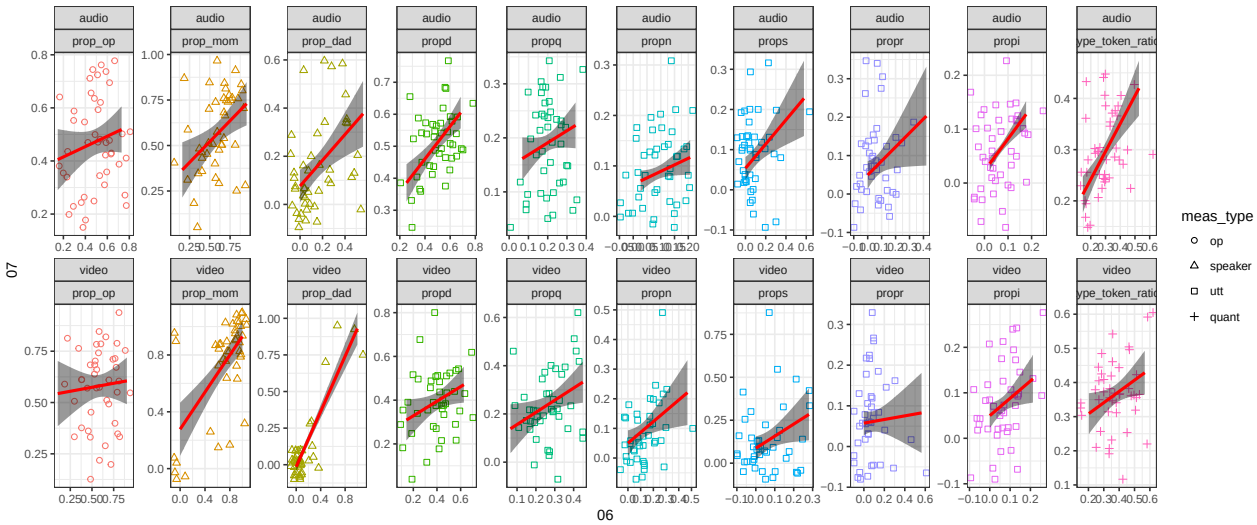

Figure 6

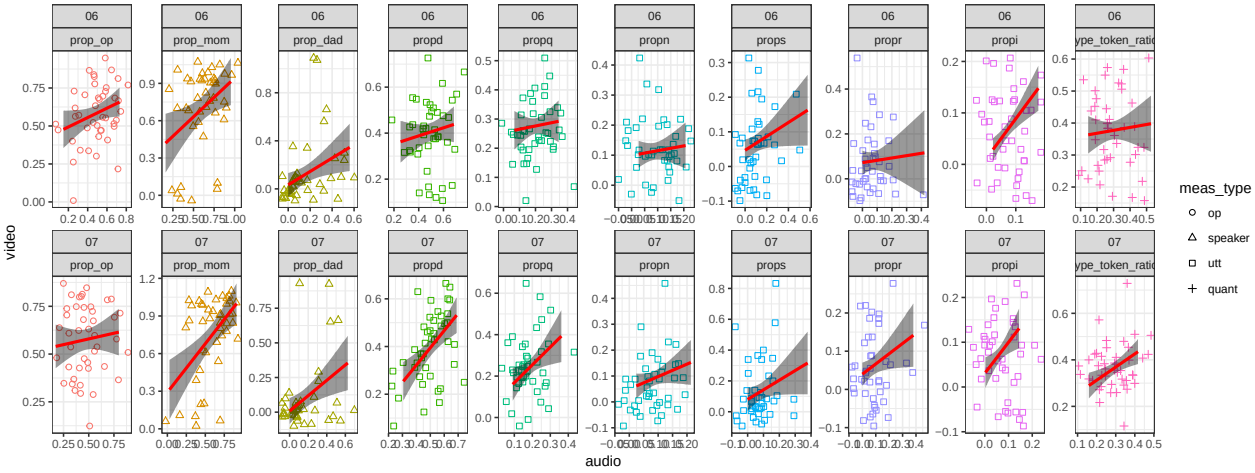

Figure 7

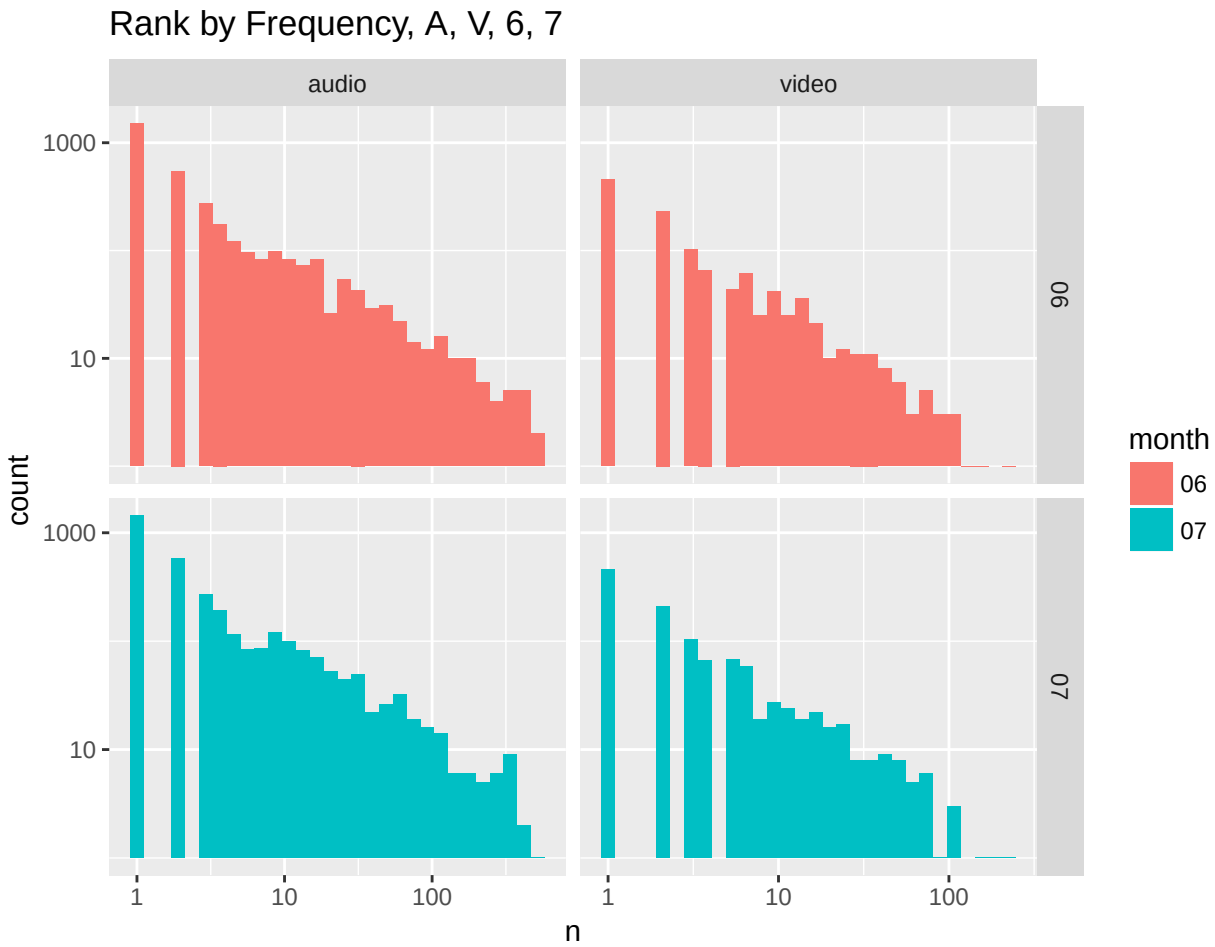

Figure 8. Zipfian word frequency distributions

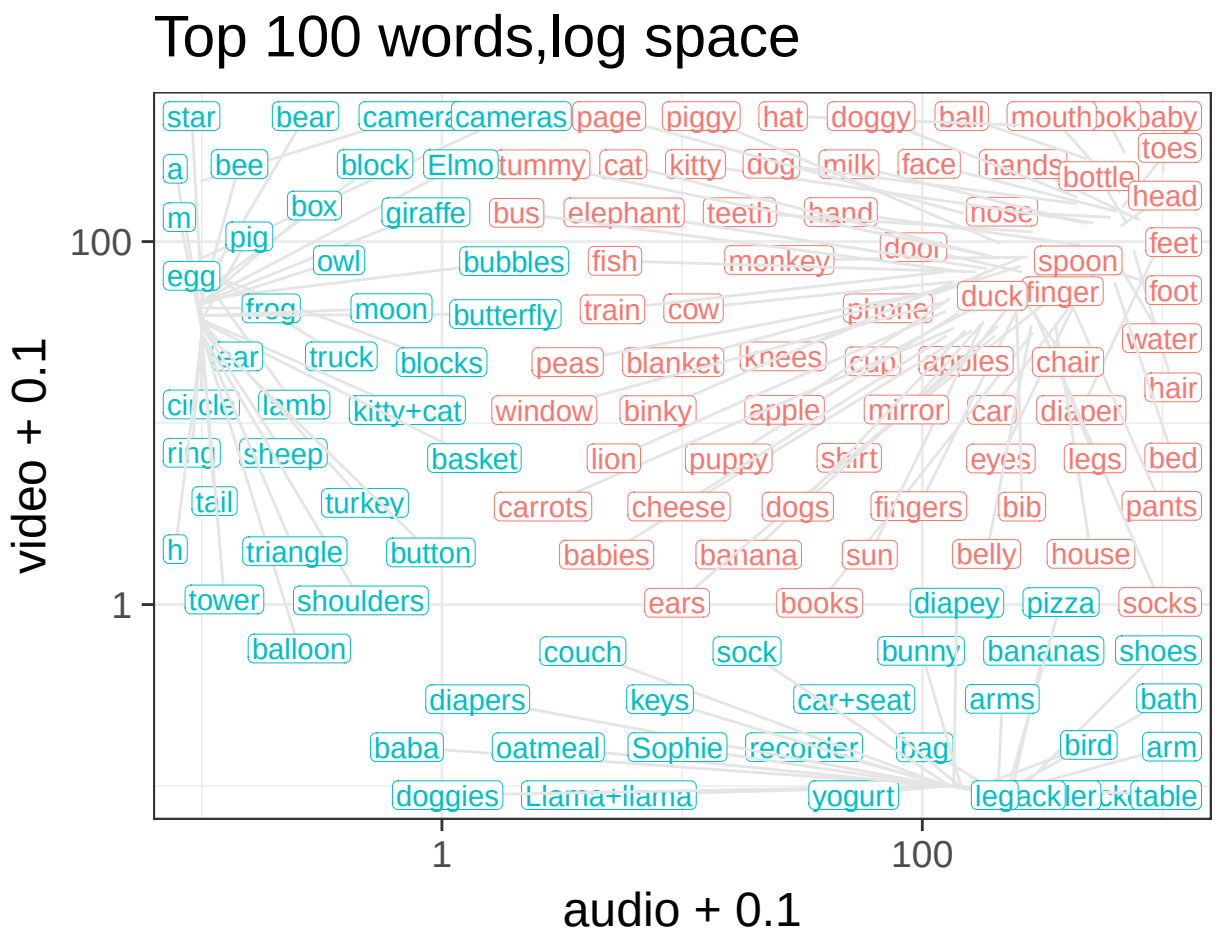

Figure 9. Top 100 words in log space

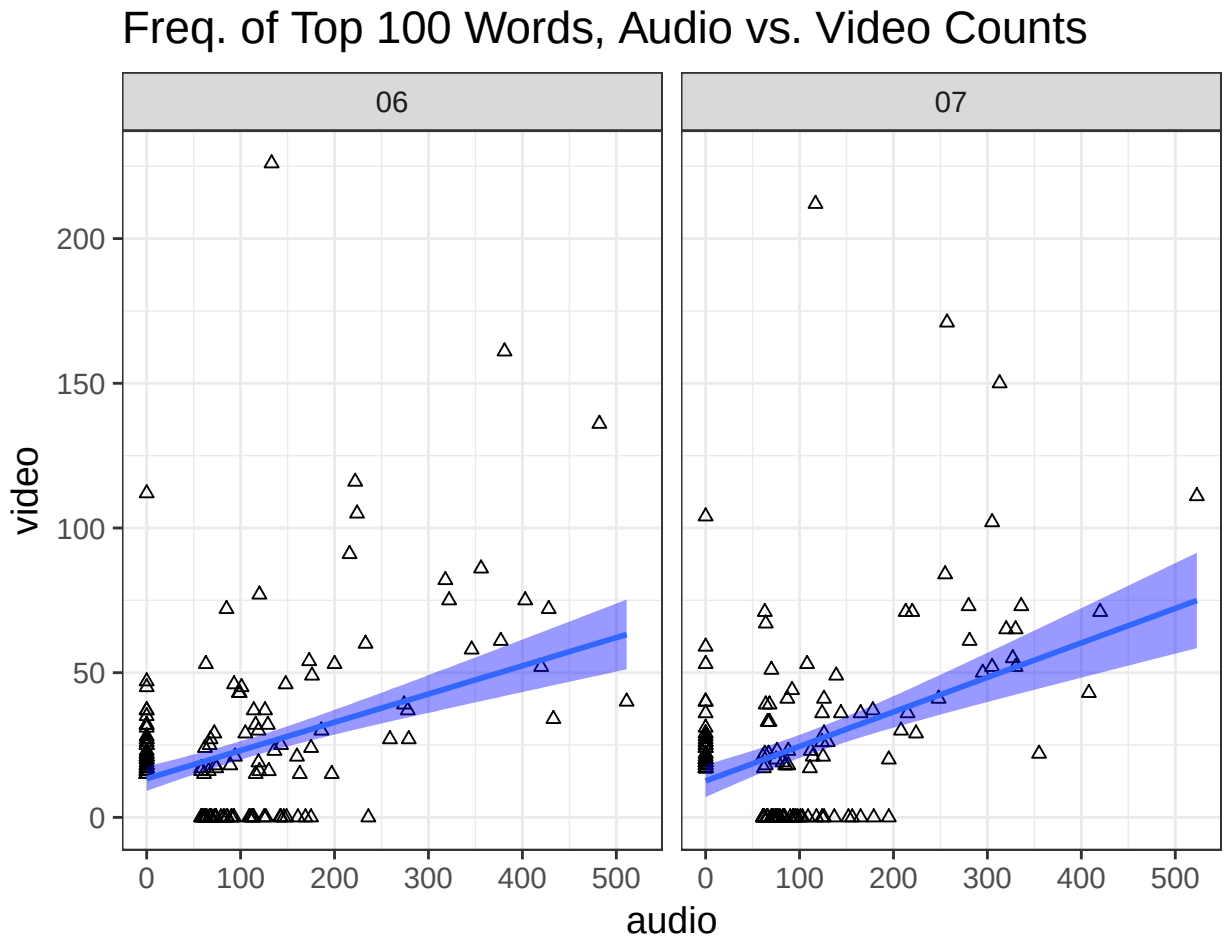

Figure 10. Top 100 words correlations by recording type

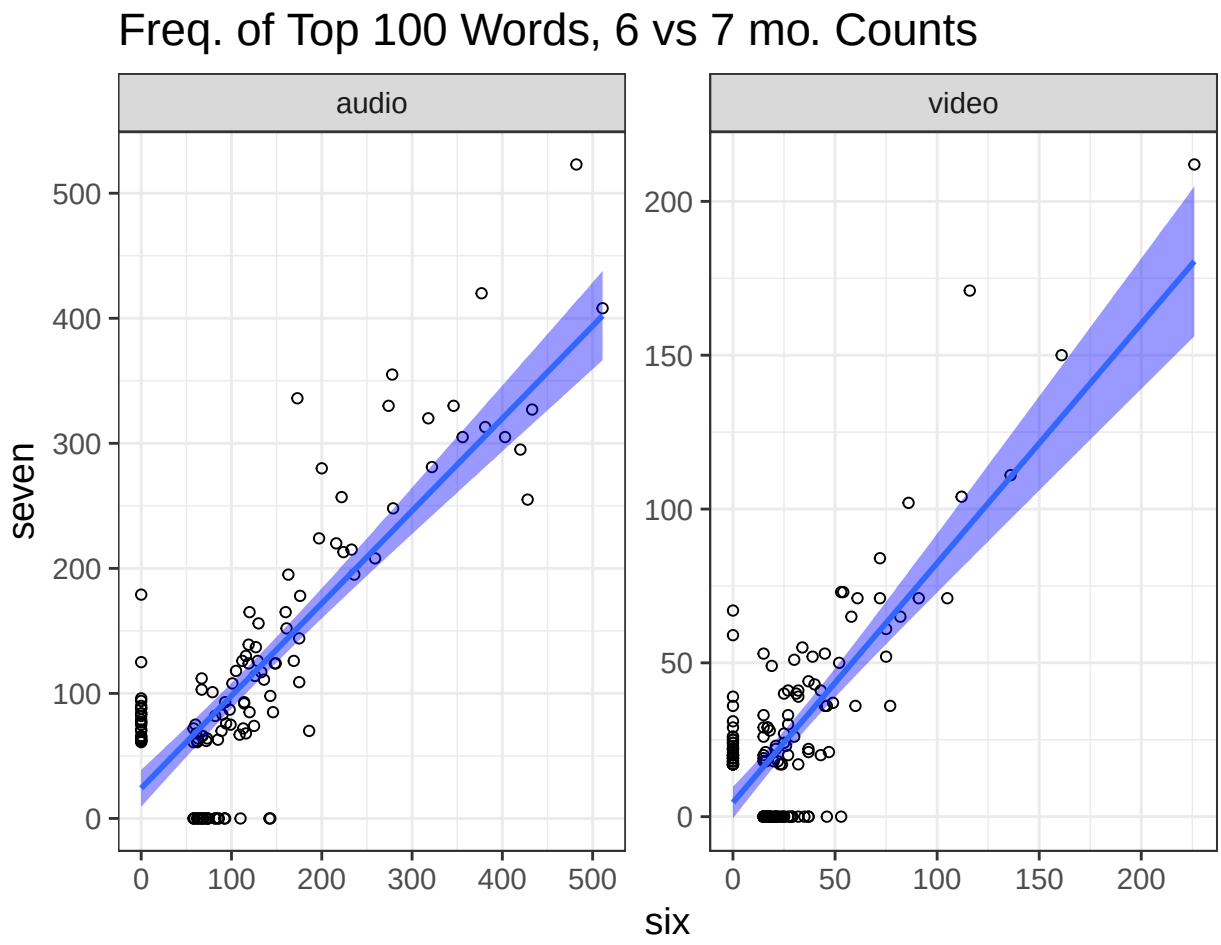

Figure 11. Top 100 words correlations by month

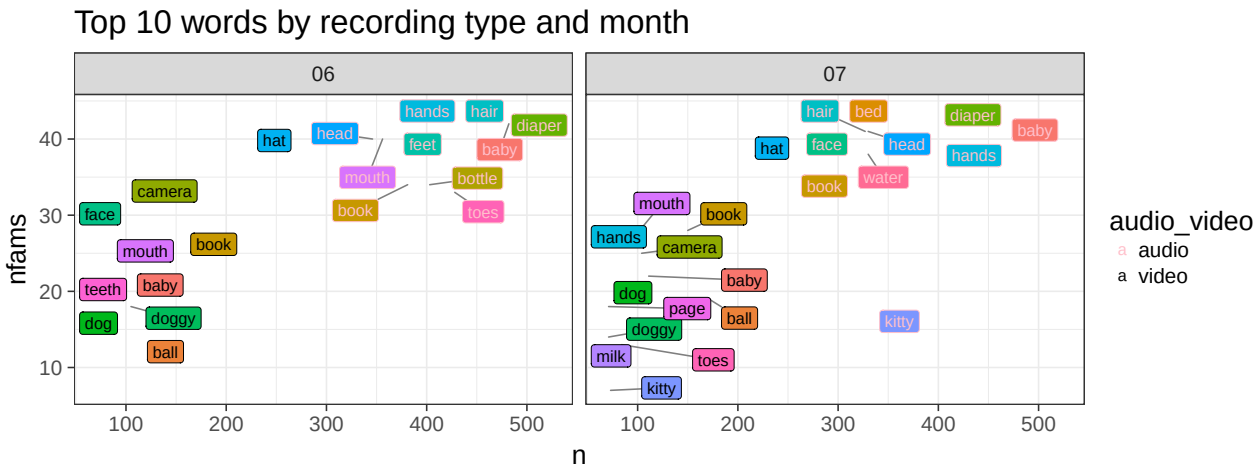

Figure 12. Top 10 words by month and recording type

Count/Prop	Month	A/V	Corr	Wilcoxon
Count	*6 vs 7*	*Audio*	*12/12*	*0/12 differed*
Count	*6 vs 7*	*Video*	*11/12 (not reading)*	*0/12 differed*
Count	*6*	*A vs. V*	*3/12 (d/q/i)*	*9/12 differed (not fat, r, s)* Type, tokens, op, q,d,i,n,,#spkr, mot)
Count	*7*	*A vs. V*	*11/12 (not s)* Type, tokens, op, q,d,i,n,r,,#spkr, mot, fat	*9/12 differed (not fat, r, s)* Type, tokens, op, q,d,i,n,,#spkr, mot
Prop	6 vs 7	Audio	7/10 (not op, q, n)	0/10 differed
Prop	6 vs 7	Video	2/10 (mot, fat)	0/10 differed
Prop	6	A vs. V	1/10 (i)	4/10 differed (not i,n,r,s, mot, fat) ttr,op, q, d
Prop	7	A vs. V	4/10 (fat, d, r, mot)	6/10 differed (not i,n,r,s) ttr,op, q,d, mot, fat

Figure 13