**Talk, You’re On Camera! Or, Comparing Naturalistic Audio and Video Recordings**

Bergelson, Amatuni, Dailey, Tor, and Koorathota

**Highlights (up to 4, no longer than 25 words)**

* We measured 44 infants’ early noun input during free-form, unobserved infant-caregiver interactions, in hour-long videos and daylong audio-recordings, at 6 and 7 months
* Infants heard 2-4 times more concrete nouns per hour during video-recording than during audio-recording
* *Proportional* measures of utterance-type, talker, and referential transparency were relatively similar in daylong audio and hour-long video recordings. Month-to-month language measures were highly stable
* There are trade-offs between observational recording techniques: the language infants hear varies in short videos and long audio-recordings, even *sans* experimenter

**Abstract (250 words)**

**Introduction**

Researchers have studied early human development by observing infants experiencing their natural habitats with their caretakers for over a century (Taine, 1876). Indeed by 1936, over 60 such ‘baby biographies’ were already published (Williams, 1936). While observational data was once limited to the shorthand notes that researchers could jot down in the moment, the advent of cheaper and less obtrusive recording technology has made it increasingly feasible to observe children in their homes, childcare centers, or in the lab, and then later annotate and transcribe these recordings for different questions of interest (XX.) Thanks to a developmental research community committed to sharing such data beyond its original purpose (e.g. Databrary and CHILDES), such recordings and annotations have led to thousands of articles about early development, and in particular, language acquisition (CHILDES, XX).

Complementarily, certain research questions within the domain of language development have proven amenable to short, controlled, in-lab studies. Such infant studies use short experiments to assess whether a group of infants of a certain age can, for instance, discriminate certain speech-sounds, learn a set of transitional probabilities between syllables in a variety of contexts, or evince comprehension of a set of words (XX).

But for questions concerning the actual linguistic, social, and cognitive input infants are exposed to in their day-to-day life, naturalistic observational data are necessary. Such data can in principle reveal what infants—individually, or on average—actually learn *from* as they make use of their biological endowments and environmental resources.

In recent years, technological advances have made it feasible to collect longer and higher-quality audio and video recordings of infants and caretakers going about their day-to-day lives, permitting richer language analyses (XX). Through such recordings, researchers have been able to better describe the normative range of variability, both across typically-developing individuals and populations, and longitudinally, within them (XX). These more ecologically-valid recordings have also permitted deeper dives into the differences in “language nutrition” in high-and low-socioeconomic status homes (XX), and into the properties of the language environment in the context of clinical diagnoses such as maternal depression, autism-spectrum disorders, premature birth status, and ADHD (XX).

Depending on the goals of a study, there are several key decision-points even within the realm of high-quality naturalistic recordings. One concerns which modalities to capture: auditory, visual, or both; there are dovetailing tradeoffs to each. Audio recording equipment tends to be unobtrusive, and has long battery life, but audio alone can make understanding the context more challenging. In contrast, video recordings tend to be richer, providing data on gestures and facial expressions, as well as the visual scene. However, video recorders generally have a shorter battery life, and are more obtrusive to participants, especially if head-mounted.

Another decision when collecting observational data concerns whether the experimenter should be present during recordings (e.g. in the home or a lab ‘observation booth’). A third decision-point concerns whether to capture structured interactions, or free-ranging ones. Further decisions concern whom to sample, when, and for how long.

Researchers often operate as though any path through such seemingly mundane decision-trees leads to equivalent results, yet this is rarely tested directly. This problematically creates a body of research whose theoretical conclusions are built on equivalency assumptions that go unmeasured.

In one recent study comparing two paths down such a decision-tree, Tamis-Lemonda et al. (2017) compared mother-infant behavior in both a 5-minute structured interaction, and in 45 minutes of free play. Sessions were video-recorded in the home by an experimenter, and subsequently transcribed. The results showed that relative to free play, in the structured interaction infants generally experienced far denser language, both in quantity of words (i.e. tokens) per minute and in word-variability (i.e. types) per minute. They also found that language quantity between these 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 quantities of language input from small samples must be made cautiously.

In the present study, we explore these issues further, directly comparing hour-long video recordings and daylong audio recordings in a single sample of 44 infants, once at 6 months, and once at 7 months, as part of a larger study on early noun learning. We used a ‘sparse’ annotation system, in which we annotated any time a concrete noun (generally, an object, food, pet, or body-part word) was said to the target infant, or said loudly and clearly in the child’s presence, along with three further properties:

1. Type of Utterance (declarative, question, imperative, short-phrase, singing, or reading)
2. Object-referent Co-presence, i.e. was it a referentially transparent situation (yes or no)
3. Talker (including live interlocutors and electronics)

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, relative and absolute 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), but also, along with their ‘absolute’ counterparts (type counts and token counts), predict language development (XX). For others metrics, we simply don’t 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. We ask: how robust are linguistically-relevant measures 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? We predict that infants’ daily experiences are relatively consistent at six and seven months, and if that is what we are capturing in our recordings, we expect to find strong convergence across time-points. Several accounts are compatible with cross-month differences, though given such a result, it would be impossible to rule out the possibility that caregivers simply behavior differently at initial versus subsequent home visits.

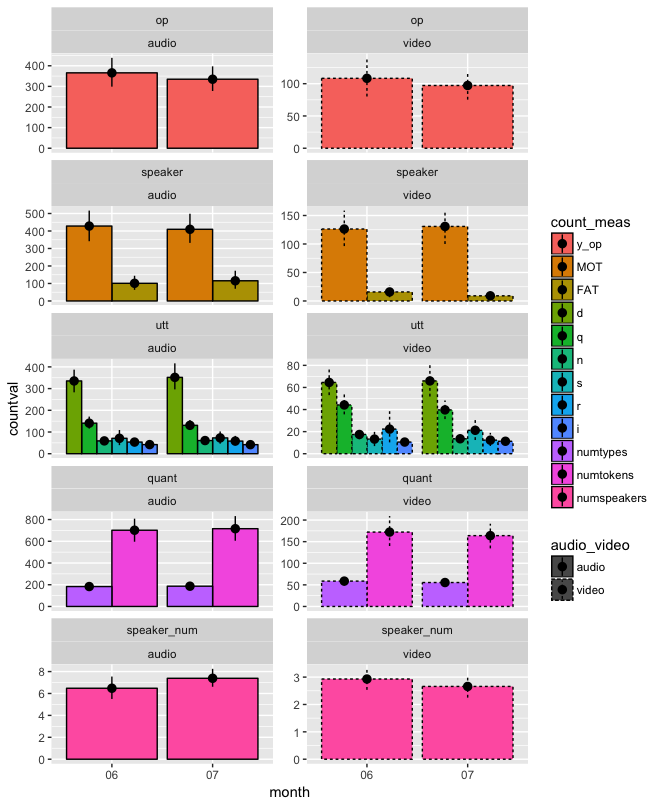
Thus, our main goal was to compare the language input young infants receive, as measured by an hour of video and a (separate) full-day audio-recording, each at two time-points. This seemingly methodological question has deep implications for developmental theory: we examine how sampling and aggregation approaches may limit (or expand!) our conclusions about the linguistic input that in turn drives early cognitive development.

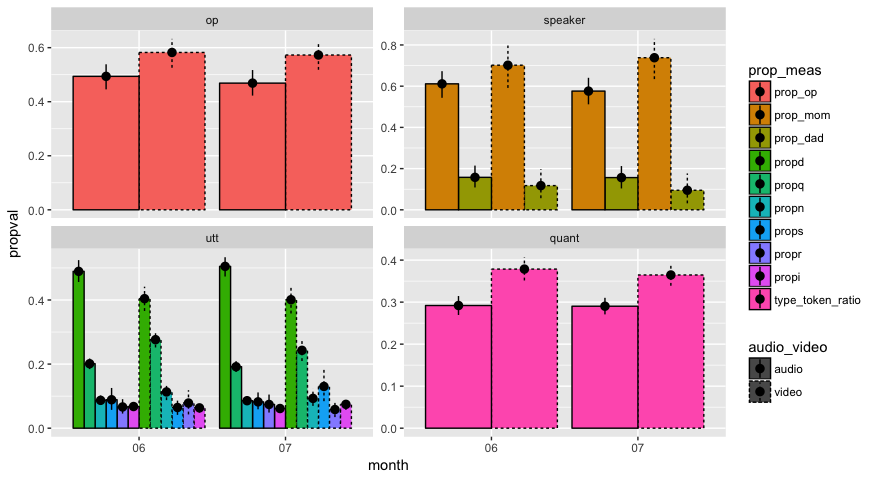
**Methods**

**[ask staff to fill in]**

**Results**

|  |  |  |
| --- | --- | --- |
| **Measure** | **Derived Count** | **Derived Proportion** |
| Quantity | Noun tokens, Noun types | Moving Avg. Noun  Type-Token Ratio |
| Speaker | Nouns from Mother Nouns from Father  Unique Speakers | Prop. Nouns by Mother  Prop. Nouns by Father |
| Utterance Type | Nouns in Declarative, Imperative, Question, Short-Phrase, Reading, and Singing Utterances | Prop. Nouns in Declarative, Imperative, Question, Short-Phrase, Reading, and Singing Utterances |
| Object Co-presence | Nouns said when the referent was present and attended to | Prop. Nouns said when the referent was present and attended to |



Based on the coding scheme above, we derived count (n=12) and proportional (n=10) measures from the annotations in each file. See table XX. We then aggregated the data at the recording level, generating derived count and proportion values for each child, for each recording-type, for each month. For all recording type and month comparisons, we look at whether the data *varied* significantly (by two-tailed, paired Wilcoxon-Test), and *correlated* significantly (by Kendall Rank Correlation) across the given groups. We applied Holm’s p-value adjustment for multiple comparisons (Holm, 1979).

**Derived Counts, Month 6 vs. 7**

We first analyzed the 12 derived count measures in month six versus seven, by recording-type. Across children, all12 derived count audio measures correlated significantly month-to-month (Kendall’s tau ranged from .26 to .51; all adjusted-p<.05), and did not significantly differ by month (all adjusted-p>.05). All video measures too correlated significantly month-to-month (Kendall’s tau ranged from .23-.60; all adjusted-p<.05), except number of reading tokens (adjusted-p>.05). For video too no derived count measures varied significantly from six to seven months (all 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 of the 24 correlations). This suggests that parents are likely behaving relatively naturally during recordings, or are adjusting their behavior as a function of our observations to an extremely consistent degree.

**Derived Counts, Audio vs. Video-recordings**

We next assessed how quantity of input scaled between our hour-long video recordings and daylong audio recordings. While video recordings were all around an hour (Mean, SD, Range, XX), the length of audio recordings varied (M=14.4(XX sd), R=10.5-16 hrs). The modal recording length was 16 hours, the maximum capacity of the LENA devices. However, children varied in their naptime schedule, and indeed from audio alone it can be hard to know with certainty whether the child was asleep. As described in the Methods section, we demarcated all long stretches of silence and thus have an upper-limit on their wakeful (i.e. non-silent) hours (Mean, SD, Range.). This is in line with established norms for 6–8-month-olds in the US (Mandel et al, 2010), which report the average amount of sleep is 3 hours during the day, and 10 hour during the night. Thus, given a 24-hour day, 16 hours of recording beginning when the child wakes up should capture roughly 11 daytime hours, in line with our silence demarcations.

To examine how the video data ‘scale’ to day-length data, we divided the audio-recording measures by the number of wakeful hours in each audio-file, providing an hourly average that could be compared to the video-recording data. We then divide the quantity from the (hour-long) video-recording by the hourly audio-average to compute a ‘Video Boost’ for each metric. For example, infants heard XX noun-tokens on average in their audio recordings, and XX in their video tokens. This renders an average of XX tokens per audio hour (after dividing each audio-recording by its number of wakeful hours), and thus a XX/XX = XX Video Boost. See Table XX.

Another way to compute the difference between the daylong and hourlong recordings is to simply divide the audio count by the video count. Doing this results in a range of 2.5-4 across our metrics, rather than the tenfold increase one would expect if video recordings captured an average hour of the day.

**Still need to do stats for this section after normalizing by length**

**Derived Proportions, Month 6 vs. 7.**

Unlike the count-based measures, the 10 proportional measures from each recording-type are readily comparable without further computation. Descriptively, within recording type, the differences month-to-month were small. Indeed, as with the derived count measures, there were no significant differences between month six and seven within audio or video recordings (all adjusted-p>.05). The pattern of correlations differed somewhat from the derived counts: for the audio recordings, 7/10 proportional measures correlated from month 6 to month 7 (for these seven measures: Kendall’s tau range: .30-.46, adjusted-p<.05; for object co-presence, questions, and short-phrases adjusted-p>.05). For video data, only proportion of input from mom and from dad correlated from month 6 to month 7 (Kendall’s tau = .44 and .63, respectively; adjusted-p<.05). Thus, overall, at the group level, the proportional metrics were very stable 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.

**Derived Proportions, Audio vs. Video-recordings**

Across recording types, proportional measures differed more substantively; see figure XX. Within the speaker-based metrics, the video recordings featured mothers more and fathers less than the audio recordings (by ~15% and 5% respectively, XX). This likely reflects the demographics of our dataset: nearly all families had full-time working fathers, but mothers’ work schedules varied, and video recordings took place on weekdays during the typical work-day hours. Videos featured more object co-presence (~10%) than audio-recordings, likely reflecting the less far-ranging activities families engaged in while video equipment was present. Regarding quantity, videos also had a higher type to token ratio (i.e. more lexical diversity), even using the MATTR (Fergatiotis et al, 2011), a point we return to later.

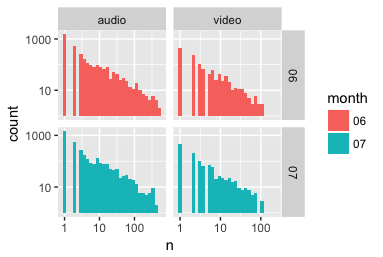
Across utterance-types, at a gross level, the distributions of nouns in audio and video-recordings were similar, with declaratives making up most of the noun input, followed by questions, with the remaining input spread relatively evenly across imperatives, reading, singing, and short phrases. However, audio recordings featured relatively more declaratives and fewer questions than videos (XX, and XX%, respectively).

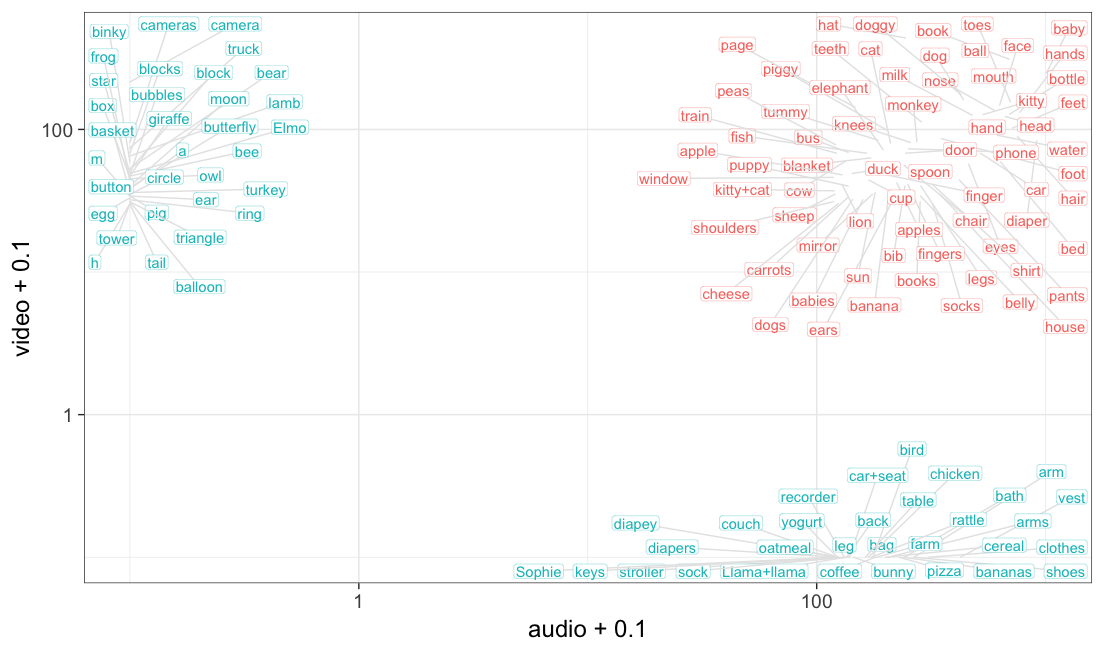
Statistically, at six months, 4/10 proportional measures differed significantly between audio and video recordings (proportion of object presence, declaratives, questions, and type-token ratio, all adj-p<.05), though the average estimated difference was only 2.4%. Only the proportion of imperatives correlated significantly across recording-types (Kendall’s tau = .31, adj-p<.05). At seven months, the same four proportional measures differed significantly across recording types (avg. estimated difference, 2.5%), in addition to the proportion of input from mothers and from fathers (all adj-p<.05). However, three of the four variables that differed significantly at seven months also correlated significantly (proportion of noun input from mom, dad, and in declaratives), along with the proportion of nouns heard in reading in the input (all adj-p<.05, Kendall’s tau range: .30-.33). We return to appropriate interpretation for this mixed pattern of correlations and differences in the discussion.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **P\_op** | **P\_mom** | **P\_dad** | **P\_d** | **P\_i** | **P\_r** | **P\_q** | **P\_s** | **P\_n** | **ttr** |
| **W\_6** | **X** |  |  | **X** |  |  | **X** |  |  | **X** |
| **W\_7** | **X** | **X** | **X** | **X** |  |  | **X** |  |  | **X** |
| **C\_6** |  |  |  |  | **X** |  |  |  |  |  |
| **C\_7** |  | **X** | **X** | **X** |  | **X** |  |  |  |  |

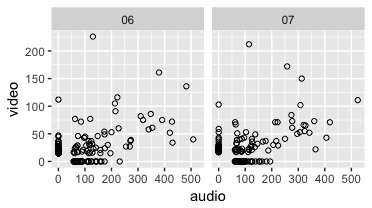
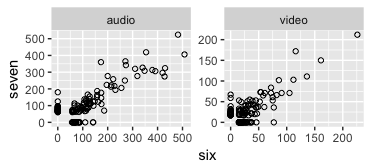
**Noun Frequency and Prevalence**

We next characterized 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 5,811 unique object words heard across months and recording types, only 3,322 were heard more than once.

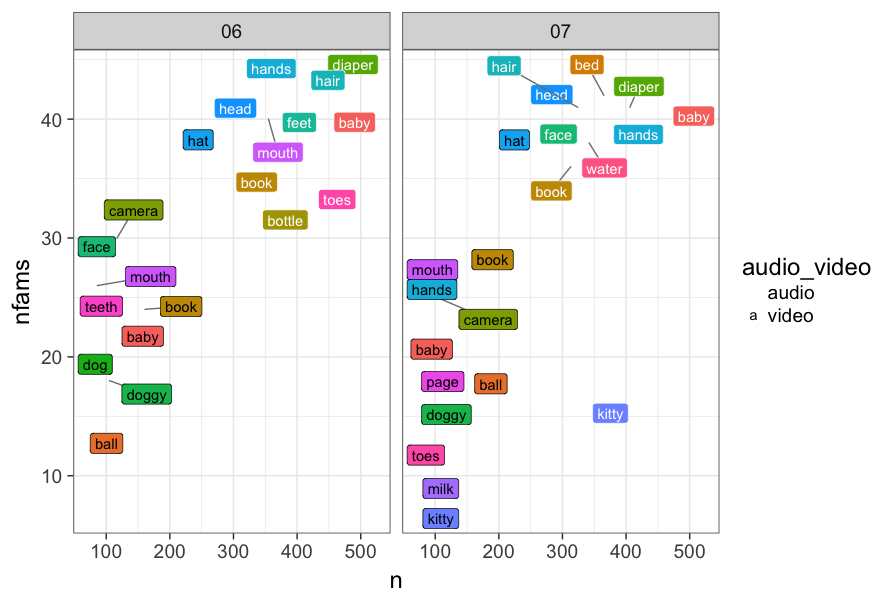




Looking at the top 100 audio and video nouns (collapsing month), we find that their frequencies across recording-types correlate significantly (Spearman’s rho=.35, p<.0001), though many these top 100 nouns occurred \*only\* in audio or video recordings; see figure XX. When we compute instead the frequency of the top 100 words by recording-type for each month, while all correlations are significant, we find stronger correlations month-to-month than across recording types within month (month 6 audio vs. video: rho = .23, month 6 audio vs. video: rho = .22, month 6 vs. 7 audio: rho = .68, month 6 vs. 7 video: rho = .36, all adj-p<.05; see figure XX).

Finally, in a descriptive analysis of the top ten words by month and recording-type, we again find greater consistency across months than recording modalities (see table XX). Indeed, top words within recording-type were largely overlapping, while only two words overlapped on all four lists (book and baby). Furthermore, if instead we characterize 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 (see figure XX).



**Discussion**

Our results can be distilled to three key findings. First, infants hear 2-4 times more concrete nouns per hour when they and their caretakers are video-recorded than when they are audio-recorded. Second, for both absolute and relative metrics of linguistic input, month-to-month differences were minimal, while recording-type differences (even accounting for recording length) were large. Finally, while both recording-types resulted in zipfian distributions of word frequencies (where many words were only heard one time), the top words within hourlong video-recordings were far less consistent across families than top words within daylong audio-recordings.

***[implications for point 1]***

***[implications for point 2]***

***[implications for point 3]***

Understanding what infants learn *from* is a key part in understanding what and how they learn at all. For instance, as we alluded to above, one critical question within language development concerns understanding how infants aggregate the linguistic data they are exposed to. That is, for any important input property, does it matter that infants have a high *proportion* of it, a high *quantity* of it, or both? And to what degree are these metrics related? Here we have taken first steps in understanding how two different data collection approaches may influence the answer to this question. Without knowing how our sampling methods may be limiting us in principle, we necessarily limit our ability to adequately model infant learning.

Admittedly, performance or “Hawthorne” effects are possible for both hour-long and daylong samples, but as noted by Suskind and colleagues regarding interventions, “sustaining increased talk for a 10-hr recording day is much less likely than being on best behavior during [a] 1-hr videotaped session...”. Anecdotally, researchers find that during daylong audio recordings, families generally slip back into a ‘natural’ mode of interaction within half an hour of turning on the recorder (Weisleder, *personal communication*).

[...]

One limitation of the present study is that there may be a fairly large degree of self-selection into this study: some families are unwilling to invite researchers into their homes to listen and watch their interactions, even when the observer isn’t present during recording. The present results suggest that within families who opt into participating in such a study, caretakers modulate their behavior as a function of recording length and modality.

Relatedly, while it’s clear that infants heard more language per hour during video recordings than during audio recordings, it’s difficult to assess whether the videos represent the same ‘kind’ of input, just more densely, or whether they’re better seen as a ‘best behavior’ scenario for how caregivers think they should act. Based on the proportional results, we have some evidence in favor of both[...]

More subjectively, the kinds of everyday interactions we capture in daylong audio recordings (bathroom events, family members rushing to get out the door or get dinner on the table, sibling quarrels, etc.) tend to ‘feel’ more natural, in ways that our metrics here may do not capture. However, it’s impossible to say whether this is due to the recording length or modality; we suspect it is both, though the development of less obtrusive video recorders with better battery capacity may prove otherwise.

**Conclusions**