Article:

Laing, Catherine orcid.org/0000-0001-8022-2655, Kalenkovich, Evgenii, Koorathota, Sharath et al. (10 more authors) (Accepted: 2025) A Year of Nouns From English-learning Infants' Daily Lives: the SEEDLingS-Nouns. Behavior research methods. ISSN: 1554-351X (In Press)

https://doi.org/10.31234/osf.io/3a487_v1

A Year of Nouns From English-learning Infants' Daily Lives: the SEEDLingS-Nouns Dataset

Evgenii Kalenkovich[1], Sharath Koorathota[2], Shaelise Tor[2], Andrei Amatuni[3], Shannon Egan-Dailey[4], Charlotte Moore[5], Catherine Laing[6], Hallie Garrison[4], Gladys Baudet[4], Federica Bulgarelli[7], Sarp Uner[4], Lillianna Righter[1], & Elika Bergelson[1]

[1] Harvard University
[2] University of Rochester
[3] University of Texas-Austin
[4] Duke University
[5] Concordia University
[6] University of York
[7] University at Buffalo

Author Note

Other than first and last authors, authors are listed in the order that they joined the project.

Correspondence concerning this article should be addressed to Elika Bergelson, 33 Kirkland Street, Cambridge, MA 02138, USA. E-mail: elika_bergelson@fas.harvard.edu

Abstract

This paper describes a dataset consisting of manually annotated nouns from a corpus of longitudinal day-long audio and hour-long video recordings collected monthly from 44 babies from age 6 months to age 17 months. This dataset was created as part of a larger project, called SEEDLingS, that examines the development of infants' language comprehension before and after their first birthday, from earliest comprehension to the early days of word produc tion. This paper provides an overview of the corpus, describes how and why the nouns from the corpus were annotated, and discusses considerations for reuse of this dataset for future work. The described annotations and relevant metadata are publicly available alongside this manuscript.

Keywords: corpus, language acquisition, infancy, home recordings

Word count: 9803

A Year of Nouns From English-learning Infants' Daily Lives: the SEEDLingS-Nouns Dataset

Introduction

Language is an essential component of human cognition, around which many aspects of human behavior, interaction, and communication are built. Studying language development in infants and children offers valuable insights into the inner workings of the human mind and the way language connects to cognitive and social development (Miller, 1990; Rebuschat, Meurers, & McEnery, 2017). By examining the language children experience in concert with their own evolving language abilities, we stand to gain a deeper understanding of both the nature of children's linguistic knowledge and the learning processes that support it. What follows, most centrally, is a description of a dataset of nouns heard and said by monolingual English-learning infants that may provide some insight into early language development. We see the loftier goal of the enterprise as ultimately shedding light on the complex processes that underlie human cognition and behavior.

Why begin with nouns? As articulated by many researchers over many decades (Babineau, Carvalho, Trueswell, & Christophe, 2021; Bates et al., 1994; Cartwright & Brent, 1997; Gentner, 1982; Gillette, Gleitman, Gleitman, & Lederer, 1999; Waxman et al., 2013), nouns, in most of the world's languages, offer a particularly helpful on-ramp to the child learner, and thus to the researcher interested in their learning. Across many languages (including English as in the dataset described herein), there is a hefty and persistent noun 'bias' in the first words learned by children (Bornstein et al., 2004; Fenson et al., 1994). This holds both in the productive lexicon and in the receptive one that precedes it (Benedict, 1979; Huttenlocher, 1974). Building on previous work that found early word comprehension beginning around 6–9 months (e.g. Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999), the SEEDLingS project (Study of Environmental Effects on

Developing Linguistic Skills) was especially focused on understanding what drives a word to enter the early receptive vocabulary, leading to its focus on concrete, imageable nouns.

This focus aligns with substantial cross-linguistic evidence showing noun dominance in early lexical acquisition (Bornstein et al., 2004; Braginsky, Yurovsky, Marchman, & Frank, 2019; Coffey, Zeitlin, Crawford, & Snedeker, 2024; Fenson et al., 1994). While it is not yet clear why nouns may hold a privileged position in early learning across many languages, factors like imageability, conceptual simplicity, and frequent occurrence in highly informative contexts likely play a role (Coffey et al., 2024; Gentner, 1982). Contextual factors too play a role, alongside linguistic ones. For instance, prior work finds that beyond overall cross-linguistic differences in noun use in speech to children, certain contexts elicit more nouns from caretakers playing with their toddlers than others (e.g. book-sharing vs. toy play) (Choi, 2000). Admittedly, it is easier to query knowledge of nouns than other parts of speech using currently available methods (Casey, Potter, Lew-Williams, & Wojcik, 2023; Meylan & Bergelson, 2021; Wojcik, Zettersten, & Benitez, 2022), and notably, some languages are an exception to this cross-linguistic tendency (Casillas, Foushee, Méndez Girón, Polian, & Brown, 2024). All of that said, a noun bias in early lexical development remains a robust finding across most studied languages, making nouns a strong starting point for investigating early word learning mechanisms, especially in English.

A central component of the SEEDLingS project was to capture the language input of infants going about their everyday lives with their caregivers, and to then test their comprehension of both everyday nouns commonly heard by all children (e.g., *shoe*) and specific knowledge regarding the nouns and referents present in the home environment of each child (e.g., *kangaroo* for a stuffed animal from Australia which may be common for one but likely not most children in a U.S. sample). To do this we collected home recordings, annotated nouns and some of their properties in the recordings, and tested infants' knowledge of a subset of these nouns. This paper focuses on the annotated home recording dataset itself (hereafter SEEDLingS-nouns), though in published and ongoing work we explore many dimensions of these children's learning (e.g. Bulgarelli, Mielke, & Bergelson, 2021; Laing & Bergelson, 2020; Moore, Dailey, Garrison, Amatuni, & Bergelson, 2019).

More so than lab studies or choreographed play sessions, naturalistic, home-based, child-centered recordings provide us with ecologically-valid data that is a reasonably close reflection of babies' "real life" experiences (Bergelson, Amatuni, Dailey, Koorathota, & Tor, 2018; Roy, Frank, DeCamp, Miller, & Roy, 2015; Tamis-LeMonda, Kuchirko, Luo, Escobar, & Bornstein, 2017). Such recordings allow us to correlate what babies experience to what they know about language, and to look at potential changes over time. Likewise, we capture spontaneous productions by young children, which allow us to link the maturity of their speech to their experience at various timepoints. Collecting these recordings at different ages allows for establishing temporal precedence between properties of experience and language knowledge, which then could in principle be connected to causal mechanisms regarding language learning. As detailed below, we focused on a few key properties of the nouns based on prior research regarding contribution of word segmentation, talker variability, and contextual variability to early language learning (e.g., Jusczyk & Hohne, 1997; Rost & McMurray, 2009; Saffran, Aslin, & Newport, 1996; Swingley & Aslin, 2000).

We tagged 6 particular properties of each annotated noun (1) the noun as heard or said by the child (e.g. "mousey"); (2) its lemma or "dictionary" form (e.g. "mouse"); (3) who said it; (4) when; (5) in what kind of utterance (e.g. declarative, singing), and (6) whether coders judged that the referent was being attended to by the child. By including these properties for each noun in the dataset, we sought to capture some key information about a given word's potential learnability. We go into greater detail below, but provide the overall motivation for each property's inclusion here.

The first three properties relate to variability in the data – that is, in the words heard or said (1-2) and in who produces them (3). These properties, broadly speaking, are known to support language development. For instance, lexical diversity in the language environment predicts language outcomes over and above the overall quantity of words (Anderson, Graham, Prime, Jenkins, & Madigan, 2021; Jones & Rowland, 2017; Rowe, 2012), and hearing novel phonologically-similar words from multiple speakers has also been found to support word-learning (Quam, Knight, & Gerken, 2017; Rost & McMurray, 2009). The fourth property (when a word is said) lets us tap into questions about massing and spacing in the input, burstiness, clustering, etc. at different timescales and contexts (Barbaro & Fausey, 2022; Tamis-LeMonda et al., 2017); this too has been linked to early learning (Slone, Abney, Smith, & Yu, 2023).

The final two properties relate to the context in which the word appears in the data, in terms of both the structural and social context of the word (5), and the co-presence of the relevant referent (6). Utterance-level properties, such as whether or not the word was produced in isolation or within a question, have been proposed to support segmentation and word learning (Brent & Siskind, 2001; Luo, Masek, Alper, & Hirsh-Pasek, 2022), while more contextual aspects (also captured by (5)) like shared book-reading and singing may support later vocabulary knowledge (Franco, Suttora, Spinelli, Kozar, & Fasolo, 2022; Leech, McNally, Daly, & Corriveau, 2022). Finally, prior research has found that word-referent co-presence in the input supports early word-learning (Bergelson & Aslin, 2017; Cartmill et al., 2013), and is also a useful indication of a growing productive vocabulary when the phonological form of an infant's production is ambiguous (Vihman & McCune, 1994).

We note from the outset that this dataset builds on many decades of work on the language children experience and produce, and is particularly indebted to the efforts of Brian MacWhinney and colleagues with the CHILDES projects, and HomeBank in particular (MacWhinney, 2000; VanDam et al., 2016). While some of the features of SEEDLingS overlap with prior and parallel efforts (e.g. Rowland et al., 2018), it is unique in how much data was collected from the children, its density and focus on infancy (6–17mo.), its combination of daylong audio-recordings and hour-long videos (both from a child-centered perspective), and the manual annotations' focus on nouns.

To characterize the advance in data availability this figure provides, we depict the number of nouns in our dataset relative to the rest of the North American English subset of CHILDES (Figure 1). As the figure shows, the SEEDLingS-Nouns dataset adds a substantial amount of data on the noun input to 6–17-month-olds relative to the totality of other datasets. This however, is countered with the caveat that the information here is limited to the nouns and their properties, rather than the full richness a detailed transcription or coding of other dimensions of the data would provide; such transcriptions and annotations are proceeding as part of several other projects at a smaller scale. Similarly, these data have been used in various related projects that include smaller-scale transcription samples (e.g. Bunce et al., 2024), annotations of other aspects of the data like babble and child-directedness (Bergelson et al., 2019; e.g. Cychosz et al., 2021; Laing & Bergelson, 2020), and larger-scale speech-technology based analyses (Bergelson et al., 2023; Lavechin et al., 2022; Räsänen et al., 2019). However, the original focus was on the nouns and their properties. That is the scope to which we limit the current description in this paper.

In what follows we describe relevant aspects of the project in broad strokes: the recruitment and recording processes and the other data that was a part of the overall project (with references to detailed descriptions in prior work where relevant). We then go into greater detail regarding the sections of the audio and video files from which the SEEDLingS-nouns corpus is derived, and the manual annotations and metadata that the corpus contains. We next provide details on how to access the corpus, and a few high-level descriptive visualizations to give a flavor of its contents for the key variables created. We conclude with some discussion of uses of the dataset to date and our hopes for its future

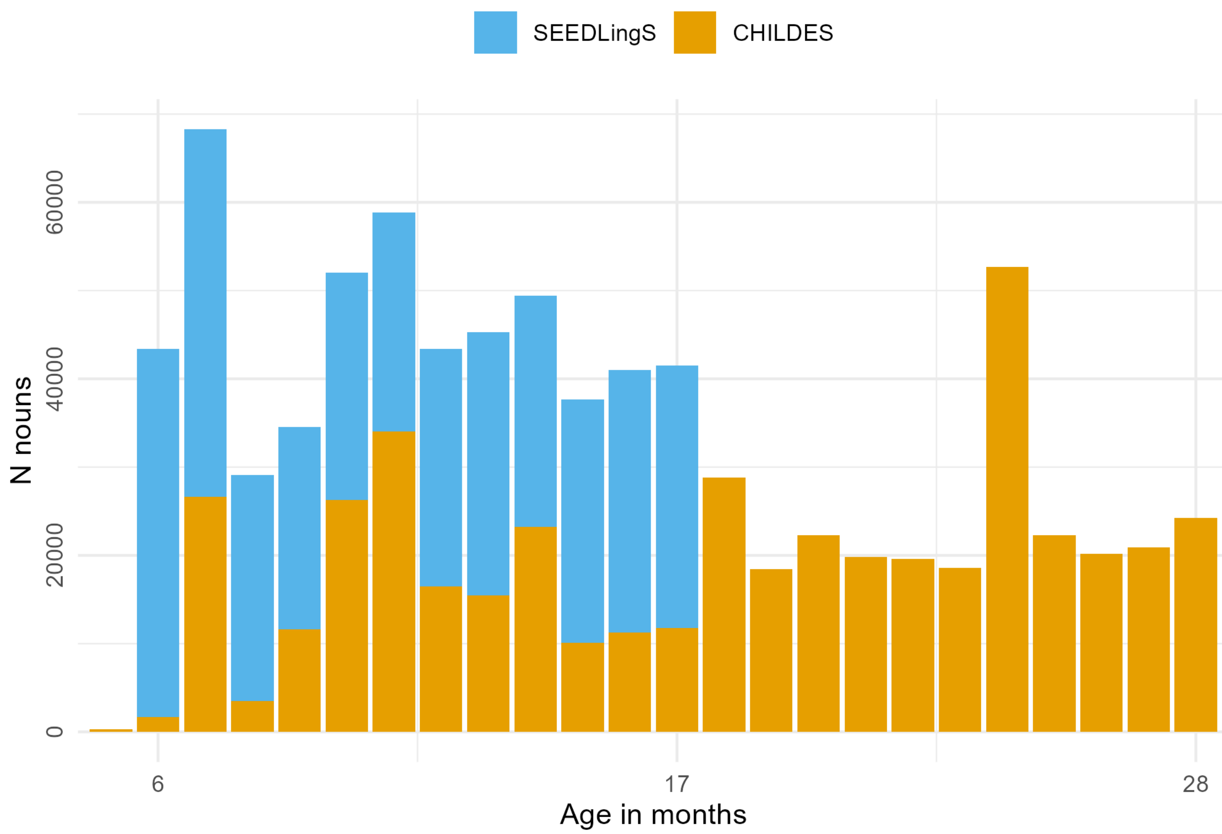

Figure 1. The number of annotated nouns produced around the target child (but not by the child) in the SEEDLingS-Nouns dataset and in the North American English portion of CHILDES, split by the age of the target child in months ranging from 5 to 28.

use by others. The amount of work that went into annotating ~360,000 noun tokens from 44 children, from ~8,000 hours of raw data (of which ~3,000 hours were annotated) was substantial, and our sincere hope is that others are able to use it to expand our scientific understanding of early language development.

Data collection

The descriptions of the participants and procedure below cover some of the same information as prior descriptions that make use of this dataset (Bergelson, 2016a; Bergelson et al., 2018; Bergelson & Aslin, 2017).

Participants

Participants (N = 44; 21 girls, 23 boys) were typically-developing, monolingual English-learning infants and their families. All infants were full-term (40±3 weeks), had no known vision or hearing problems, and heard $\geq$ 75% spoken English. Families participated longitudinally from when children were 6 to 18 months. Participants were recruited from a database of local families in Rochester, NY; all eligible families were contacted with no specific demographic-based recruitment applied. An additional two infants were enrolled and completed some study activities but withdrew shortly thereafter; their data was not analyzed or included in the sample. The final sample of 44 infants was 93% non-Hispanic White, 2% Hispanic, and 5% multiracial. In terms of parental education, 75% of mothers and 72% of fathers had at least a Bachelor's degree. Participants were compensated $340 for the yearlong study.

Procedure

Data collection began in November 2014 and ended in July 2016. Child-centered daylong audio recordings and hourlong video recordings were collected each month from 6–17 months, for a total of 12 audio- and 12 video-recordings per infant. For audio-recordings, parents were given a LENA audio recorder and shirt or vest with a specialized pocket in advance of their recording day so that recordings could begin first thing in the morning on the recording day. For video-recordings, on a different day that same week, two researchers came to the home, set up the camcorder on a tripod in the corner of the main room parents indicated they would spend time in (which they could also readily move). They further equipped the child with a hat or headband with two Looxcie cameras attached, one pointed slightly up, and one slightly down. If an infant refused the headwear during setup, researchers gave the parent a headband with a Looxcie camera instead. Researchers then left and returned after one hour. The broader study (not the present focus) included in-lab visits every other month (6–18mo., n=7), and monthly vocabulary and gross motor surveys (6–18mo., n=13 of each) (cf. Moore et al., 2019).

The audio from the LENA recorder was then processed by the LENA proprietary software (Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011) and exported as paired .cha and .wav files for manual annotation in CLAN (MacWhinney, 2000). The video feeds were combined in Vegas software into a single feed and exported for manual annotation in Datavyu (Datavyu Team, 2014).

Data annotation and aggregation

This section describes our annotation methodology, the resulting dataset structure, and the procedures used to ensure data quality. We begin by outlining the database organization, then detail the specific annotation criteria and procedures, followed by our validation approaches.

Tables derived from the annotations

The dataset is organized into 4 key derived tables: **seedlings-nouns**, **recordings**, **regions**, and **sub-recordings**. The **seedling-nouns** table serves as the main data source, containing all 358,300 annotated noun tokens and their properties. There is 1 recording of each type missing from the complete dataset: one video was accidentally deleted due to a technical error; one audio-recording was deleted upon parent request.

The **recordings** table lists each recording, its duration, and the amount of time that was listened to by annotaters. This is detailed further below but in short, all video recordings (generally one hour) were annotated in full while audio recordings were annotated for different amounts of time depending on the age of the child (6 and 7 months: full day; 8–13 months: 4 hours; 14–17 months: 3 hours). The **regions** table indicates the *regions* (i.e., spans of time) that were listened to in each audio file. Finally, the **sub-recordings** table lists the time of day the audio and video recordings were started. A small subset of recordings (n= 36) contained pauses (i.e., caretakers stopped recording by pressing a button for a period of time because of, e.g., privacy concerns); the **sub-recordings** table has entries corresponding to each sub-recording. Each table is accompanied by a codebook that contains descriptions and technical information about each column. The relationships between the tables are schematically represented in Figure 2.

Annotations

In this section, we detail the main components of the manual annotation. There is an even greater amount of detail in an accompanying supplemental GitBook wiki, which is live here and exported as an archival PDF on OSF (https://osf.io/r9pvn).

Annotation criteria: which nouns to annotate. Trained coders identified each concrete, imageable noun token that was said clearly in proximity of the target child (or by the child). To operationalize “concrete” and “imageable”, we created a set of guidelines, which at its core, sought to include nouns that one could depict in a standalone way. The methodological choices here dovetailed with the broader project goals which involved eyetracking studies where these same infants would be shown images of nouns, including a subset of those being annotated. The shorthand term for this type of concrete noun was “object”, though we did include distinct sub-parts (e.g. “foot”,“tooth”), liquids (e.g. “coffee”), animals (e.g. “cat”), etc. To maintain cross-coder consistency, we created a dictionary of commonly occurring nouns with a guide for whether to code them, and an in-house set of guiding principles; see details in the wiki linked above.

For example, we avoided coding terms referring to people (e.g. “man” or “grandma”), with some exceptions. We did code instances of “baby” when used generically (“See the baby in the book?”) but not as a term of endearment (“Hey baby, give me a minute”), people who had specific occupations (e.g. “firefighter”) and people picked out as characters in a book or show (“Batman”). We also didn’t include nouns if they were being used metaphorically (e.g. “cat got your tongue?”). The referent of the noun did not have to be

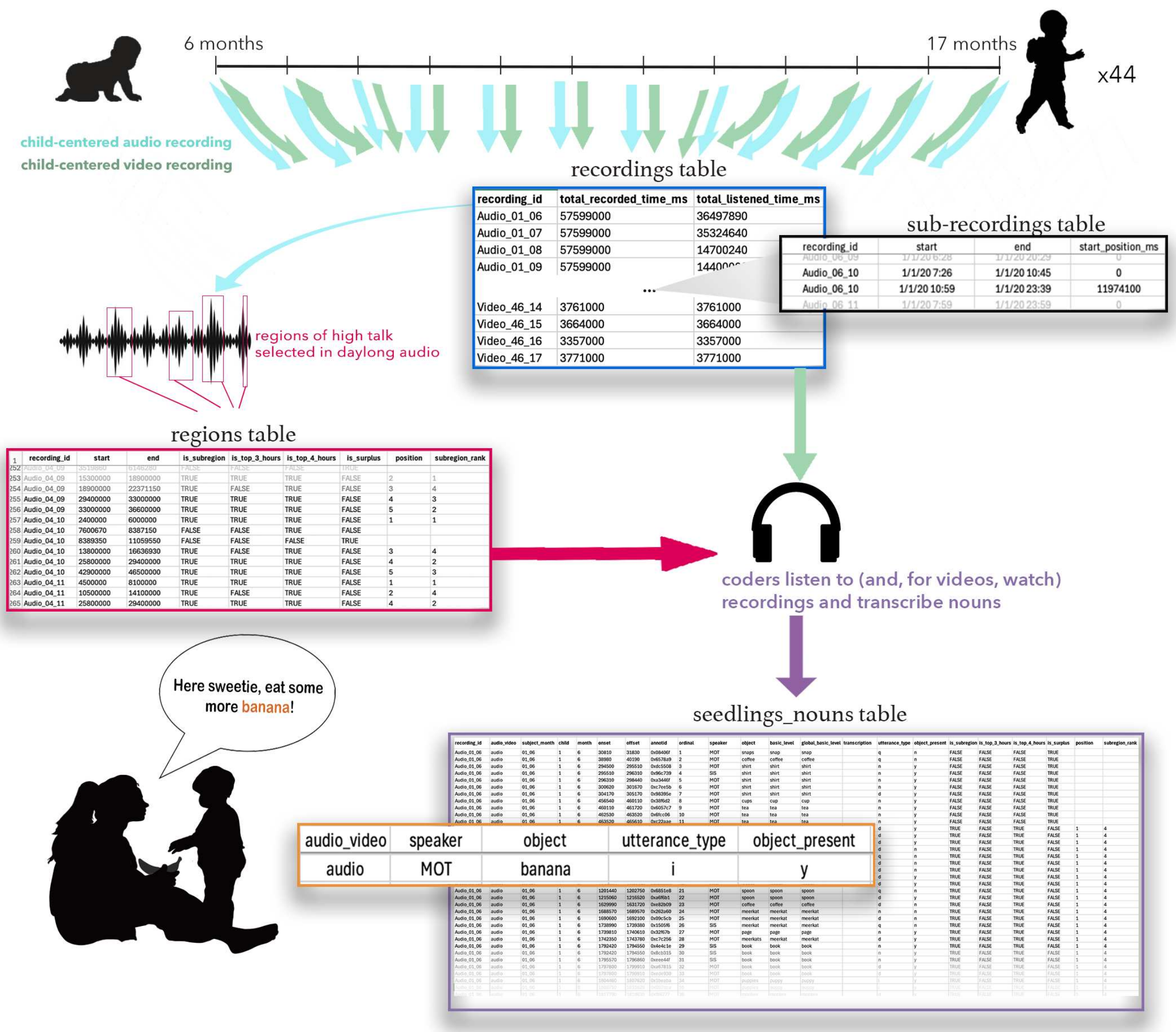

Figure 2. A day-long audio and an hour-long video recording was collected each month from 44 infants aged 6–17 months (see tables **recordings** and **sub-recordings**). Video recordings and audio recordings from month 6 and 7 were annotated in full. For months 8–17, a set of *high-talk regions* within the audio was selected for annotation (see table **regions**). All concrete imageable nouns produced by or near the child were annotated (see table **seedlings-nouns**).

visible to the child to be annotated (this was tagged separately, see "object presence" below.). In short, we annotated just the kinds of everyday common nouns that occur in the early English-learning baby's vocabulary, and refer to a concrete entity.

While some research with longform recordings has specifically looked at child-directed speech (e.g., Weisleder & Fernald, 2013), focusing specifically on speech directed to the target child, nouns in this dataset did not have to be directed to the target child wearing the recorder in order to be coded, but distant nouns in the background (e.g., television heard from the other room) were not coded. In practice, around 80% of the tokens of the most common nouns were rated as being produced in child-directed speech in a subsequent tagging effort (Bulgarelli & Bergelson, 2024).

Annotated properties. For each annotated noun token, coders noted the word as it was said, what type of utterance it was a part of (delineated below), whether the referent of the object was attended to by the child, and who produced the word. In the aggregated **seedlings-nouns** table, these correspond to columns *object*, *utterance_type*, *object_presence*, and *speaker*.

Before elaborating on these tags, we first explain the relevant notion of an utterance; further below we clarify how this relates to the timestamps we provide for each tag. Each annotated noun occurred within an utterance. Utterance was operationalized in keeping with the CHILDES guidelines proposed by MacWhinney and colleagues in the CHAT manual (MacWhinney, 2000). Adult utterances generally contain a main clause with rare exceptions for imperatives (e.g. "No, pasta!"), and given our use case, all contained at least one concrete noun. Guidelines for child utterances similarly followed CHILDES guidelines, with prosody and pauses as a guide for non-adult-like constructions (which were extremely common in the expectedly sparse productions of nouns in this dataset's age-range of 6-17mo.)

The **object word** was annotated with any relevant morphology or diminutivization (e.g., *tootheroo* for *tooth*); this tag was later expanded to include two versions of the lemma (basic level and global basic level, described under "Word lemmatization" below). The **utterance-type** tag was a closed-set decision among 6 categories (as well as a rarely-invoked unsure (`u`) option): declarative (`d`, e.g. "Yup, you found the truck!"), imperative (`i`, e.g. "Put your socks on."), non-utterance (`n`, e.g. "Yes, water.")), question (`q`, e.g. "Did you drop your plate?" , "Where's the doggy?"), reading (`r`, e.g. "The very hungry caterpillar…"), and singing (`s`, e.g. "The itsy-bitsy spider went up the water spout…"). Coders were encouraged to make their best guess among the closed-set categories; *unsure* was ultimately used for 0.002% of all tokens for utterance-type.

The first four of these (declaratives, imperatives, non-utterances, and questions) roughly correspond to the *syntactic* structure of the sentence; the (admittedly oddly-named) "non-utterance" tag was used for words in isolation, fragments, or standalone noun-phrases of fewer than 3 words; questions included both polar and wh-questions. The last two utterance-type tags reflected two common activities (reading and singing) which have a cluster of properties that tend to differentiate them from spontaneous speech (in prosody, sentence length, and lexical variety). Various combinations of these categories could overlap (e.g. a sung question); we implemented a hierarchy for such cases as follows: reading > singing > non-utterance > question > declarative, imperative, consistent with prior approaches (M. Soderstrom, personal communication, Dec. 29, 2013). For example, declaratives and questions read in a book would be tagged as reading, and a single word with rising intonation would be tagged as a non-utterance.

Object presence (i.e. whether the referent of the noun was present and attended by the child) was tagged as yes (`y`), no (`n`), as well as a rarely-invoked unsure (`u`) option. Here too, coders were encouraged to make their best guess among the closed-set categories; *unsure* was ultimately used for 0.3% of all tokens for object presence. Coders inferred object presence from context (visual and linguistic for videos; only the latter for audio-recordings). While sometimes this was very clear ("No, we don't have bananas", "Are those your sandals?"), determining object presence in audio recordings presented unique challenges relative to video recordings (since in videos referents could be directly observed most of the time.) In general, coders relied on several contextual cues to infer object presence, especially for audio-recordings: verbal descriptions from speakers (e.g., "Look at this ball"), sound effects indicating object interaction (e.g., toy squeaking, book pages turning), conversational context and dialogue continuity, and explicit references to shared attention (e.g., "Do you see this? It's a giraffe!"). Unsurprisingly, we did see somewhat lower reliability for audio-recording object presence tags, as we discuss further in "Data validation and checking" below.

Finally, each speaker was assigned a 3-letter code that reflects their relationship to the target child. This code remained consistent for a given speaker across all recordings for a given participant. Thus, a child with two aunts might have `AU1` assigned to Aunt Sally and `AU2` to Aunt Debra for all recordings from that child. Across all recordings and participants, the same code was used for the same type of relationship whenever possible. For example, `MOT` always represents the Mother and `AU1` always represents an aunt.

To give a final concrete example of how all the codes would apply to one concrete noun, a coder hearing "Those are some beautiful brushes!" would annotate "brushes" as follows: *brushes, d_y_MOT* indicating that the object *brushes* was part of a Declarative (`d`) utterance produced when the object was present and attended to (Yes, `y`), and the word was spoken by the Mother (`MOT`).

Child production transcriptions. For only the noun tokens produced by the child (speaker code `CHI`), the *transcription* column contains a phonetic transcription (n = 5,737 tokens from 42 children). This column is empty for the remaining 98.4% of tokens because they were produced by other speakers. The transcription was done using the PHONASCII (Allen, 1988, Appendix 1: UNIBET) transcription convention with the following modifications: (1) [?] was used as a glottal stop as opposed to a rising terminal, (2) IPA symbols that could be represented using ASCII and were not part of the PHONASCII convention were used as well. This transcription method was used instead of IPA to avoid Unicode encoding issues and simplify typing.

Word "lemmatization". To further characterize the wordforms found in the dataset, two additional columns were added to permit aggregation of forms that corresponded to the same object word: basic levels and global basic levels (*basic_level* and *global_basic_level* in the table). The basic level annotation reflected a *recording-specific* collapsing of words that shared a meaning and phonological content, defaulting to the most common version of the word in that recording. For example, in a recording where "paci" is used to refer to pacifiers 15 time, "pacifiers" was used 3 times, and "pacifier" is used once, the basic level for all 19 instances would be "paci" in the data corresponding to that recording. While families often stuck to a given term for an object, this sometimes drifted across recordings. In practice this meant that in some recordings (e.g. month 7 audio for child A) an object may have *paci* as its basic level, while in another recording from the same subject (e.g. month 10 video for child A or B) it would be *pacifier*, since this was done by selecting the most common form on a per-recording basis based on broader project goals and workflows.

To further streamline across recording-specific frequencies, after all recording-level spreadsheets were aggregated across the entire sample, each noun's lemma was codified at the whole-corpus level, dubbed *global_basic_level*. The form used here was generally the "dictionary" version of the word, i.e. the most conventional format. Thus, regardless of family- or recording-specific prevalence, the global basic level in the running example would be *pacifier*. Basic levels were primarily used to determine which wordforms should be used in the concurrent eyetracking studies that accompanied the creation of this dataset, and as such needed to maintain the idiosyncrasies of each family's speech. In contrast, global basic levels were designed to bundle related wordforms into larger categories to enable a more generalized consideration of the wordforms in speech to and by young children in the sample at large.[1]

Timestamp details. Each annotation is also timestamped, though the origin of these timestamps varies by recording type. For videos, the onset and offset of each utterance with a tagged noun was created manually as a cell in a Datavyu spreadsheet (Datavyu Team, 2014). For the audio-recordings, the annotators worked in CLAN-formatted `.cha` files, where each line corresponded to a time-stamp that came from the LENA algorithm (the LENA algorithm does an exhaustive pass through each LENA recording to tag into broad speaker-classes that roughly correspond to utterances; cf. Cristia et al. (2021)). Thus, for video, the onset and offset of the utterance were manually generated for each token individually, while for audio, LENA software-based automatic utterance segmentation was used.

For both video and audio, multiple noun tokens could share the same onset and offset, as these reflect the utterance borders, not the start and stop of a given token. For instance, if the utterance was ‘This book has a penguin in it!’, *book* and *penguin* would have the same onset and offset timestamps, encompassing the broader utterance. The LENA-generated “utterances” were often shorter than what a human coder counted as an utterance in the video recordings. It’s worth noting that there is a margin of error in the manually generated video timestamps due to the lag in streaming video from a VPN-protected storage server and the brief duration of many utterances, as coders set timestamps based on real-time reactions (with cross-checking of a subset in a cleanup

[1] While all tokens had a *global_basic_level* assigned to them, 6.9% of them have an NA for *basic_level*. This happened because the basic levels were originally used to select nouns most commonly heard by the child in order to use them as stimuli for the in-lab looking preference eyetracking study and certain words were not included due to either being too generic or centered on private body parts/bodily functions. For example, the 5 most common nouns based on global basic level that do not have a basic level assigned are *toy*, *food*, *poop*, *buttocks*, and *picture*.

pass). This difference in what was considered an utterance in video versus audio, along with the potential imprecision of the video timestamps, was acceptable for our purposes (and for many others): our primary objective was to confirm that the tokens were produced within the specified boundaries, rather than to pinpoint utterance boundaries with great precision. But for other use-cases, e.g. determining utterance length, one should use timestamps with caution and with these caveats in mind.

Top-3 high-talk hours. Finally, we include a boolean column *is_top_3_hours*, which indicates whether the token was produced during the top 3 most talkative hours of the audio recordings, as estimated using LENA automatic metrics (detailed further below). As this is only relevant for the audio recordings; the 30.4% of tokens that came from the videos have empty values in this column. Each noun token also has a unique identifier, referred to as an *annotation id*, abbreviated *annotid* in the **seedlings-nouns** table. Table 1 summarizes the columns of the **seedlings-nouns** table.

Data checking and reliability

An annotation effort of this scale required significant research staff effort, following specialized training. The initial data collection took place from 2014–2016, with subsequent cleaning and aggregation efforts taking place thereafter. To ensure a high level of data quality, there were multiple passes through the annotations, with the 'closed-set' categories (i.e., object presence and utterance type) receiving further reliability assessment. The initial annotations were created by trained researchers at the university where the data were collected, who were also interacting with the families directly and thus came to know them well. Each recording's annotations were first run through scripts for initial error checking and then manually checked by a staff member (ensuring closed-set vocabulary and certain spelling conventions were followed) who also added the "basic level" coding described above. Data were also spot-checked by the PI, who, for example, looked at parts of every video as part of a stimulus-selection process for the related eyetracking studies.

Table 1

Columns of the seedlings-nouns table. See text for further details.

Column	Description
Annotations	
object	The noun as it was said, e.g., *baba* (bottle), *diapers*, etc.
speaker	3-character speaker code, e.g. `MOT` and `GRM` for mother and grandma.
utterance_type	Utterance type, a 1-character code. See text for possible values.
object_present	Whether the object was present and attended to (y) or not (n).
basic_level	Modal recording-level lemma (e.g. *paci*).
global_basic_level	Corpus-level lemma.
transcription	Phonetic transcription for child-produced nouns only.
Timestamps	
onset, offset	Onset and offset times of the utterance containing the noun, in ms.
Identifiers	
annotid	Unique noun ID. A 6-digit random hex, e.g., *0xe4d823*, *0xcba546*.
audio_video	Recording media: *audio* or *video*.
subject	2-digit participant code from *01-46* with *05* and *24* skipped.
month	2-digit month code from *06-17*.
subject_month	Combination of *subject* and *month*, e.g., *23_07*, *03_12*.
recording_id	Combination of *audio_video*, *subject*, and *month*, e.g., *audio_23_07*.
ordinal	Noun order in a recording; distinguishes nouns in the same utterance.
region_id	Region ID in the **regions** table. NA for videos.
sub_recording_id	Sub-recording id in the **sub-recordings** table.
Other	
is_top_3_hours	Is this noun/utterance in the top-3 high talk audio hours? `TRUE`/`FALSE`.

Note. The descriptions here are adapted versions of the ones provided in the "seedlings-nouns_codebook.csv" file.

After this initial pass of annotation that was conducted as the data were collected, there was a second pass that checked each annotation. Finally, to establish reliability of the coding for the closed-set categories (utterance type and object presence), 10% of each file

Table 2

Reliability of close-set annotations (Cohen's kappa), split by recording-type (audio, video) and tag (utterance type and object presence). These were derived from a recoded 10% of annotations for each cell below.

	Utterance type		Object presence	
Month	Audio	Video	Audio	Video
06	0.81	0.85	0.54	0.81
07	0.81	0.73	0.59	0.64
08	0.80	0.75	0.46	0.68
09	0.85	0.82	0.55	0.63
10	0.81	0.82	0.61	0.64
11	0.83	0.86	0.60	0.73
12	0.79	0.77	0.60	0.55
13	0.82	0.81	0.63	0.61
14	0.80	0.85	0.66	0.64
15	0.86	0.84	0.52	0.64
16	0.87	0.85	0.62	0.70
17	0.83	0.86	0.60	0.64
Mean (SD)	**0.82 (0.02)**	**0.82 (0.04)**	**0.58 (0.05)**	**0.66 (0.07)**

was extracted and recoded *de novo*. Reliability (as assessed by Cohen's kappa (κ)) was generally strong, and is reported in Table 2 for each category, month, and recording type ($M_{\text{utterance\_type\_}\kappa} = 0.8$, $M_{\text{object\_presence\_}\kappa} = 0.6$.)

We do want to call particular attention to the reliability for the object presence category and note that coders were advised to use their best judgment rather than default to an Unsure rating (used for less than <.5% of tokens). This (along with the challenges of inferring object presence with only audio, as noted above) was likely part of the reason for the more moderate reliability for audio object presence coding ($\kappa = 0.58$) than we saw for video ($\kappa = 0.66$). Researchers using this variable should consider this limitation, particularly for analyses that heavily rely on object presence in audio recordings. That said, for many purposes the relatively large sample size allows for meaningful analysis of patterns even with the level of measurement noise found here; we'd welcome further refinement of this tag by interested parties.

We also conducted reliability on the child production phonetic transcriptions. To so do, a random 10% subset of all child productions were transcribed by a second coder. Discrepancies were resolved through discussion. Given that infant articulatory control is still developing at 6–17 months, these early productions naturally lack some of the precision found in adult or older child speech. To account for this, the reliability assessment focused exclusively on consonant agreement, collapsing some distinctive categories (e.g. treating all alveolar or palatal fricatives as [s]). This approach yielded an agreement rate of 77% (Cohen's $\kappa = 0.72$) for the 10–11 month data (Laing & Bergelson, 2020), a result comparable to other annotation efforts in the field. We therefore note that these phonetic transcriptions provide a reliable reflection of the broad phonetic characteristics of the infants' productions, while acknowledging the inherent developmental imprecision.

Finally, for the speaker tags, a staff member with detailed knowledge of the families (due to being part of the initial data collection and annotation staff) checked and streamlined all speaker codes for consistency. At every step of this process, data checking and correction occurred as errors were identified.

Linking between the tables

As noted above, audio and video recordings were collected from each infant monthly from 6 to 17 months, within a week of the day that the child turned a month older. We used numerical identifiers to distinguish children (*subject id* in column *subject*): 1 to 46 with 5 and 24 skipped because the corresponding participants dropped out of the study soon after enrollment.

We refer to the sequential recordings from the same infant by a *month id* (column *month*): 6 to 17. The dataset (unlike this text) uses leading zeros for both the month and subject id. For example, a token might come from the audio of infant *07* from month *08*. To uniquely identify recordings, we need to know media type, the subject, and the month id, e.g., *audio_27_12*, *video_41_03*, etc. This identifier, *recording_id*, is present in all tables and constitutes the primary key (a unique identifier) for the **recordings** table. Columns *region_id* and *sub_recording_id* are the primary keys for the **regions** and **sub-recordings** tables, respectively. All three columns (*recording_id*, *region_id*, and *sub_recording_id*) are present in the **seedlings-nouns** table so that the annotated tokens can be matched to the information in the other three tables.[2]

How much data is annotated in each file

For both the audio and video recordings, our goal was to annotate nouns from the same amount of time from the sample at a given age, as detailed below. This procedure

[2] Certain columns in the dataset are redundant with respect to other tables and share name with them. For example, **seedlings-nouns** and **regions** redundantly contain *is_top_3_hours*, and columns `subject`, `month`, and `audio_video` in all tables are redundant with these columns in the `recordings` table. This needs to be accounted for when joining the tables.

leads to a corpus with a great amount of variability across families rather than a set amount of transcribed speech per family. This approach lets us better capture variation and base rates of amount and type of speech rather than equate amount of talk; i.e. we captured all the concrete nouns in the sampled (time-based) sections of recordings rather than e.g. all the concrete nouns in 100 utterances/recording.

Audio annotation considerations. All video recordings were fully annotated with the coding scheme described above. For audio files, we were not able to annotate the entire daylong file due to practical limitations: stimuli for the eyetracking experiments (which occurred every other month, (i.e., 6, 8, 10…) relied on the previous two months' annotations, and the study had staggered enrollment over a 9 month period for a 12 month study; it simply wasn't possible to fully annotate 44 ~16-hour audio files and 1-hour video each month, even with 3–4 full time staff and a dozen RAs working 10 hours/week. Another consideration was that large periods of time in the recordings contained no speech, generally while children slept. These were marked automatically by a silence-finder in Audacity, then visually inspected (with border adjustments as needed) by lab staff, and skipped during annotation. We will refer to these regions as *silent* or *silences* below. Finally, given infants' advancing language production skills and the reduction in naps over the first two postnatal years, there was an increasing amount of speech over 6–17 months in the daylong audio files as sleep patterns shifted.

Together, these considerations led us to seek an approach that identified sections of the file for annotation that would maximize our annotation capacity but also permit standardized comparison over developmental time, and across recording-types (audio vs. video). We next detail how we selected subregions of audio to annotate such that there was a minimum of 3 annotated hours per recording. The final result is different sampling densities (full-day for 6-7 months, 4 hours for 8-13 months, 3 hours for 14-17 months), but demarcation permitting easy selection of the same amount of time in each recording when needed for a given use case (cf. Bergelson et al. (2018) for relevant analysis and discussion of subsampling).

Selection of *high talk* audio regions. We considered several automatically-generated metrics from the LENA software to home in on high-talk areas of the recordings: conversational turn count (CTC), adult word count (AWC), and child vocalization count (CVC). We opted to prioritize CTC and CVC, as these would ensure the child was awake and vocalizing, and most likely interacting with others. We thus operationalized talkativeness as the average of CTC and CVC, calculated it for hour-long running windows with a 5-minute step, and then sequentially (avoiding overlap) selected the windows with the highest value of this combined metric. In total, 5 hour-long *subregions* were programmatically delineated in this manner. For months 8–13, the four highest-ranked of these hour-long *subregions* were then annotated; for months 14–17, the top three were annotated. The remaining *subregions* were used as "backup", that is, in case there were problems with others (e.g., a clear indication the child was asleep despite the high talk ranking, as indicated by snoring or parent commentary).

Makeup, *extra*, *surplus*. In order to produce a comparable duration of annotations across all recordings of a given age, we aimed to delineate three or four hours of annotated time, within a 15 minute margin for each recording. To that end, several adjustments to the high talk sampling were necessary. For instance, the *silent* regions described above sometimes overlapped with the hour-long regions prioritized for annotation (e.g., if a parent chatted with their child while feeding them a snack and then read 3 books, this would likely be a very high-talk hour even if the child then slept for the last 20 minutes of it as part of their afternoon nap). In other cases, the *subregions* marked for annotation contained portions that weren't completely silent but couldn't be meaningfully annotated (e.g., nothing but a muffled radio in the background) and thus were manually marked as *skips*. Whenever this happened, the annotator calculated how much time was missing and created what we called *makeup* regions of the same length within the backup *subregions* (i.e., the *subregions* ranked 4 and 5). If that still wasn't enough to get to the target amount of annotation time, annotators added time past the ends of some of the annotated *subregions* creating *extra* regions.

Occasionally, this manual process over-shot the amount of intended annotation time by more than 15 minutes. In order for it to be possible to select a consistent amount of annotated time from each recording without deleting annotations (e.g., the top 3 hours from each audio file), relevant portions of *makeup*/*extra* regions were marked as *surplus*.

Total annotated time

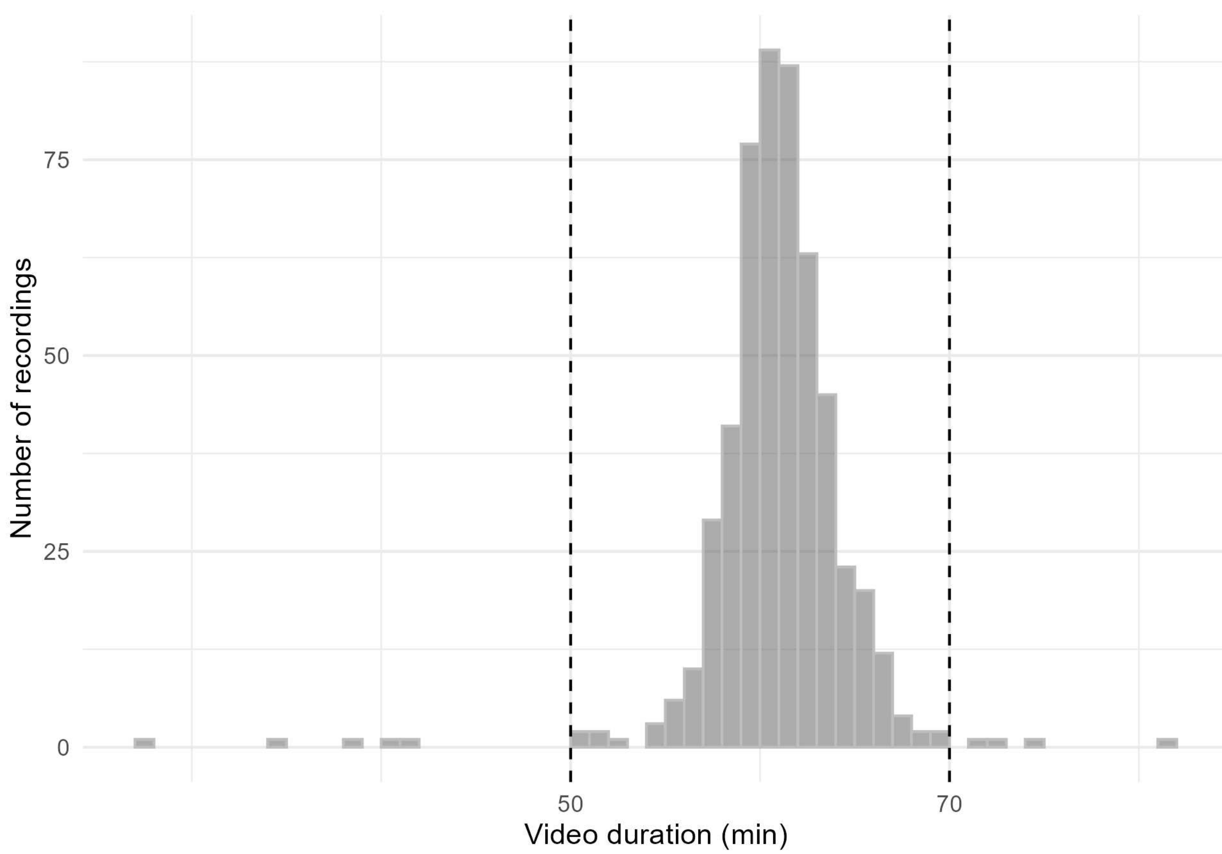

Figure 3. Distribution of video recording durations

Because video recordings were annotated in full, the amount of annotated time is

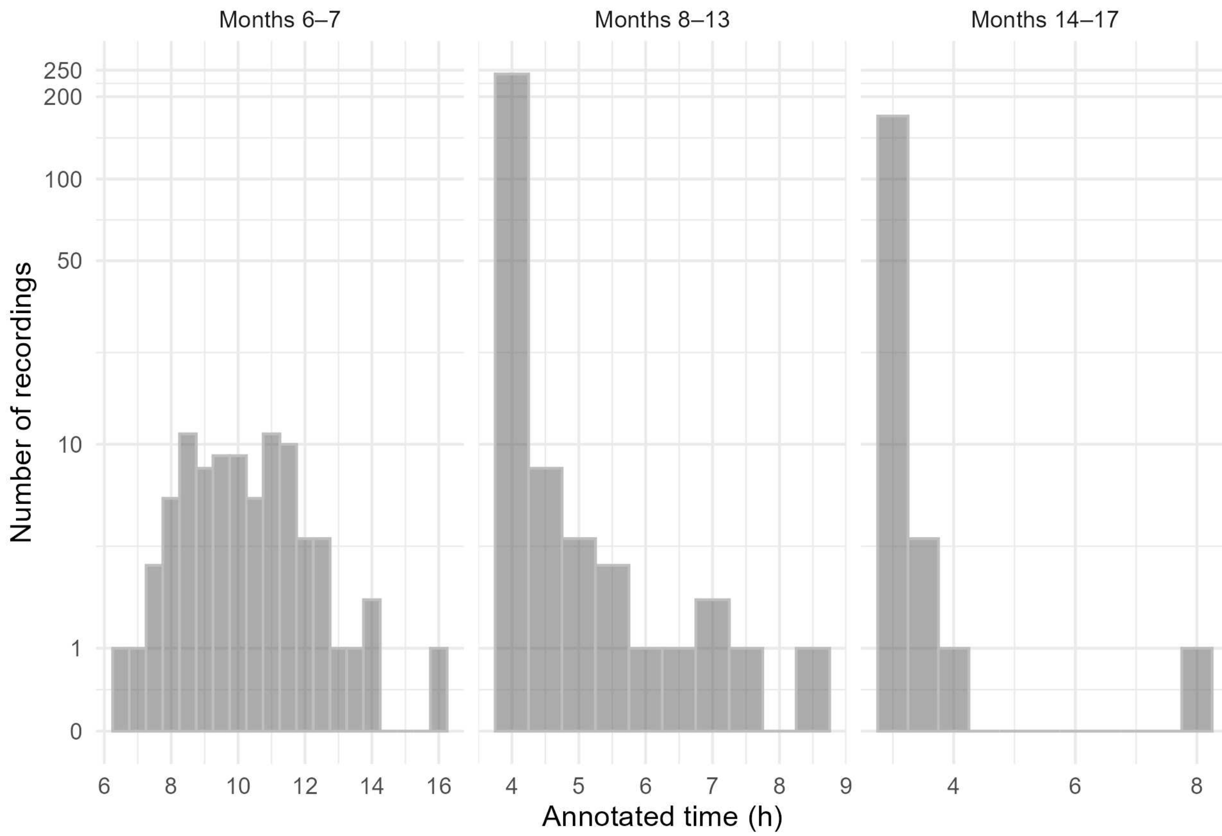

Figure 4. Distribution of the duration of annotated time in audio recordings. For months 6–7, the annotated time corresponds to the total duration of the recordings minus the total duration of silent regions. For months 8–13 and 14–17, the planned annotated time was 4 and 3 hours, respectively. However, a subset of recordings was annotated for more than the planned amount of time, resulting in *surplus* time included in the total here (and delineated separately in the dataset.

equal to the recording duration. Recordings were modally one hour, with the vast majority of recordings (518 out of 527) lasting between 50 and 70 minutes (M(SD)=60.93(3.79) minutes.) See Figure 3 for video duration and Figure 4 for the distribution of the duration of annotated time in audio recordings.

Top-X audio hours. For analyses that rely on having a consistent length of speech sampled from each family, age, or recording (e.g., Bergelson et al. (2018)), we identify tokens corresponding to the top-3 (for all audio recordings) and top-4 (for months 6–13) most talkative hours in each file.[3] For the details of the procedure we used to count annotated regions to top 3 or 4 hours, see the accompanying wiki (https://seedlings-nouns.bergelsonlab.com) or its pdf version (https://osf.io/mp9fb).

Accessing the corpus

The dataset is available on GitHub and on Zenodo as a set of CSV files. As the dataset is updated, new versions will be released on GitHub and mirrored on Zenodo. The list of published versions can be found under "Releases" on GitHub and under "Versions" on Zenodo. Changes introduced in consecutive versions are documented in the "CHANGELOG.md" file in the dataset repository.

Each data table has a companion codebook with column format information to use when loading the table. We recommend specifying data type for all columns when loading them. At a minimum, a `string` or a `character` data type should be specified for columns `child` and `month` which contain two-digit strings with leading zeros for numbers below 10.[4]

For R users, another option is to use the Bergelson Lab's internal R package `blabr`.

[3] For top-3, we added an *is_top_3_hours* column directly to the **seedlings-nouns** table to facilitate comparisons across all audio recordings. The **regions** table also includes information allowing the identification of the top x hours for other values of x.

[4] Treating these columns as numeric leads to issues downstream, as all members of our lab can attest.

Users will need to (1) clone the GitHub repository to `~/BLAB_DATA` and (2) install `blabr` from Github, e.g. with the following code:

```
remotes::install('bergelsonlab/blabr')
```

Users can then load a specific version (recommended) or the latest available version of the **seedlings-nouns** table using the `get_seedlings_nouns()` function. When doing the latter, a warning will be issued, which includes the latest version tag which we recommend users copy to the `get_seedlings_nouns()` call to ensure version consistency.

```
sn_version <- 'v2.0.0'
seedlings_nouns <- blabr::get_seedlings_nouns(version = sn_version)
# Or get_seedlings_nouns() to get the latest version (not recommended)
```

Companion functions `get_seedlings_nouns_extra()` and `get_seedlings_nouns_codebook()` allow users to load one of the three remaining tables (*recordings*, *regions*, *sub-recordings*) and the codebooks for all four. For example:

```
sub_recordings <-
  blabr::get_seedlings_nouns_extra(
    version = sn_version,
    table = "sub-recordings")
seedlings_nouns_codebook <-
  blabr::get_seedlings_nouns_codebook(
    version = sn_version,
    table = "seedlings-nouns")
```

High-level descriptive visualizations

To give a flavor of the dataset, we provide a few descriptive figures. For the descriptive analyses presented here, we specifically used the “top-3” hours subset for all

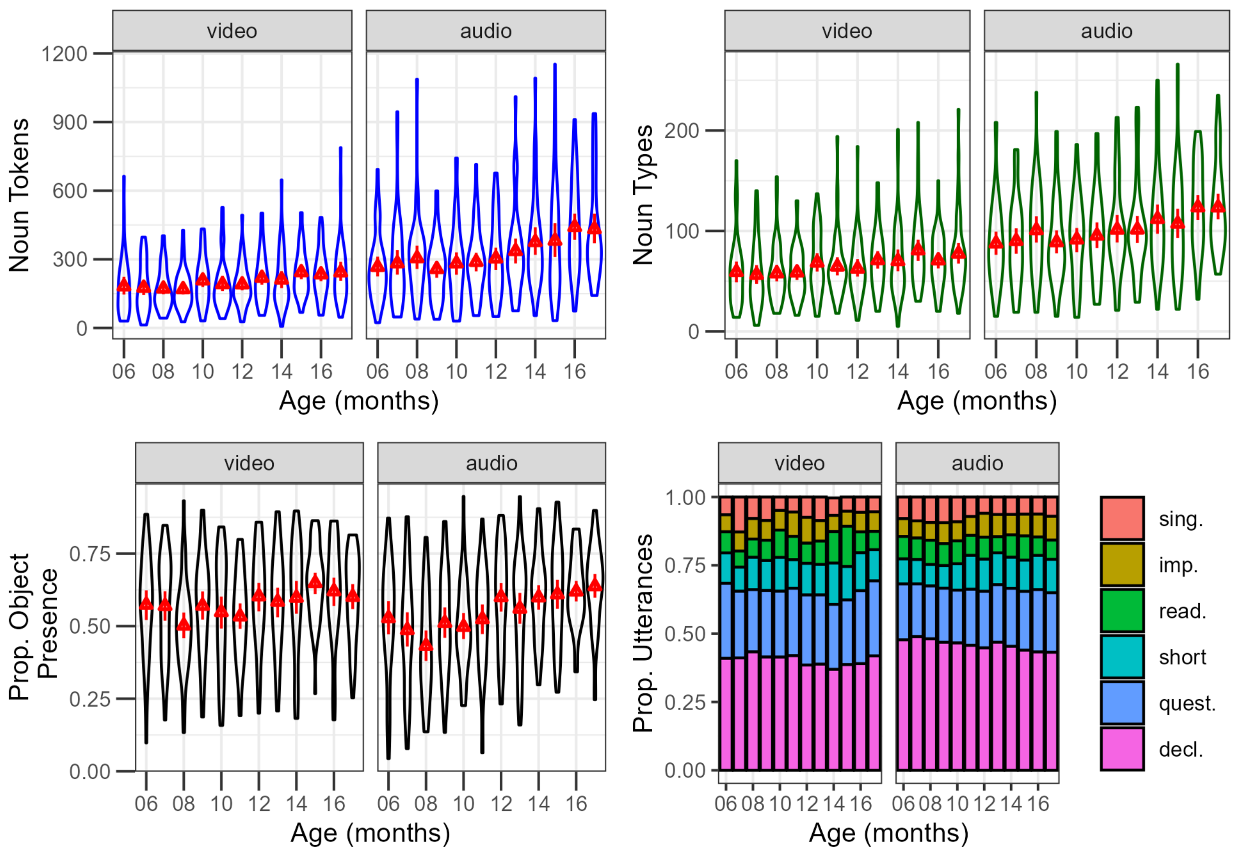

Figure 5. Noun input (tokens, types, object presence, and utterance types) by recording-type and month to 44 English-learning infants from monthly 1hr. home videos and 3hr. sampled from daylong audio-recordings. Top row: total words (tokens; blue violin plots), unique words (types; green violin plots). Bottom row: proportions of object presence (black violin plots), and utterance types (colored bars indicated the proportion of nouns heard in (sing)ing, (imp)erative, (read)ing, (short) phrase, (quest)ion, and (dec)larative utterances). In the first 3 panels, the red triangle and error bar indicate M + 95% bootstrapped CI; violin width reflects amount of data in that portion of the distribution, i.e. variability across infants in a given month and recording-type.

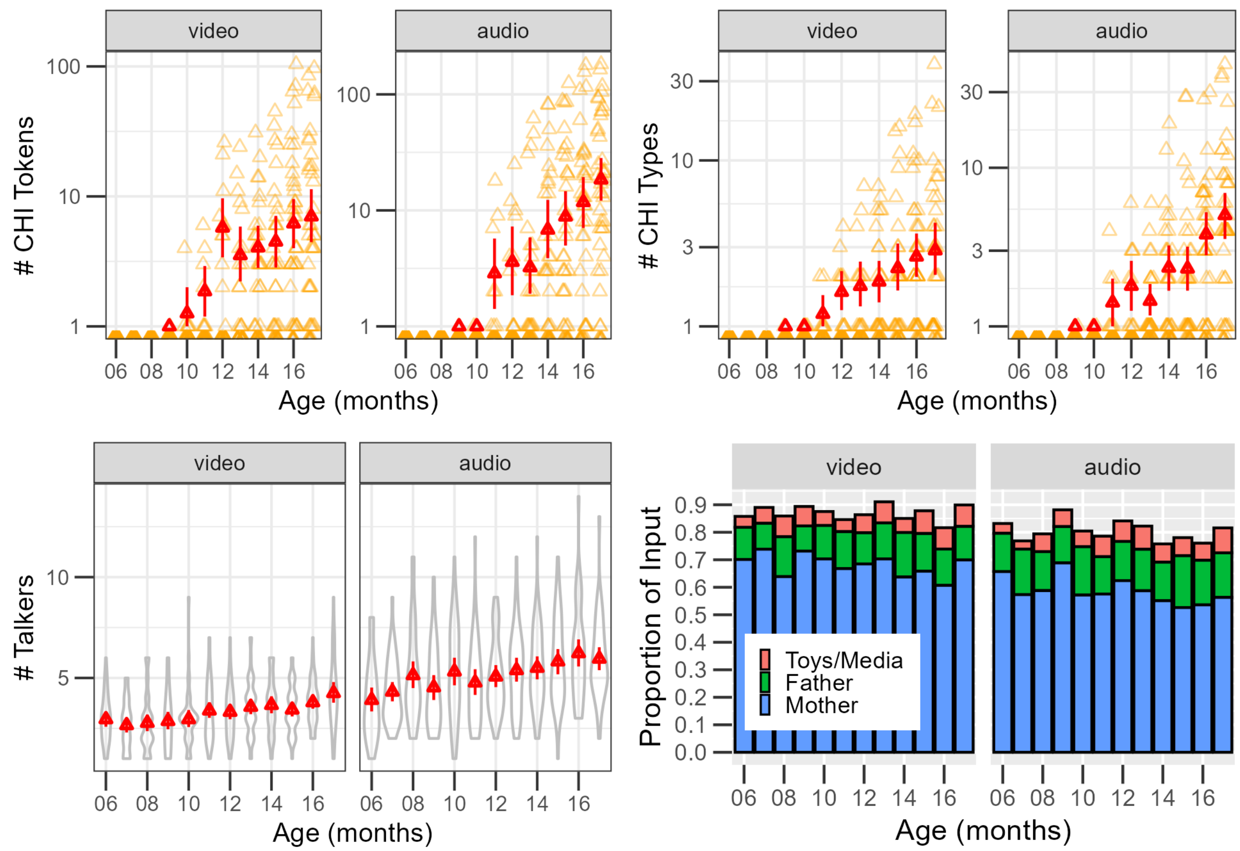

Figure 6. Child noun productions and talker distributions by recording-type and month in 44 English-learning infants' monthly 1 hr. home videos and 3 hours sampled from daylong audio-recordings. Top row: total words (tokens; left) and unique words (types; right) produced by each child (orange triangles). Bottom row: number of talkers (including the child; left), and the proportion of noun input (excluding the child) coming from mothers, fathers, and toys or media (right). Red triangles, error bars, and violins as in preceding figure.

audio recordings to facilitate comparability across ages in terms of the amount of speech coded and how it was sampled.

The first (Figure 5) shows the quantity of noun tokens and types for each month and recording type, the proportion of object presence, and the relative distribution of utterance types. To facilitate comparison across audio and video, we depict the data from the full video and from the top 3 hours of the audio as described above. We highlight 3 features of the data for readers. First, as noted in prior work (Bergelson et al., 2018), the videos have a much higher density of speech than the audio recordings, i.e. the noun tokens and type counts are 1–2x higher for audio than for video despite coming from 3x the data (i.e. 3 "high talk" audio hours vs. 1 video hour). Second, there is a large amount of variance within the recordings across children each month (i.e. the violins showing the data distribution are tall and skinny). Finally, the shifts over age are relatively subtle: there are some shifts month to month that seem to reflect noise, while others increase slightly over time. For instance, the relative proportions of difference utterance types are largely stable: declaratives and questions make up over 2/3 of the noun input every month. In contrast, the number of types and tokens creeps up with age; this is not because of the child talking (because the child is omitted from these figures) but rather, appears to be best explained by parents talking more to talkers, as explored elsewhere (Dailey & Bergelson, 2023).

The next figure (Figure 6) zooms in more on the child productions and talker data. Namely, it shows noun types and tokens produced by each child each month, the number of total talkers (including the child), and the relative proportions of noun input from key talkers (excluding the child), i.e.from mothers, fathers, and toys or media (e.g. singing toys or shows). We again highlight a few features for readers. First, we see the expected trends regarding the onset of lexical production: a period of no production followed by a huge uptick coupled with huge variability across children. By 17 months children were producing roughly 10 tokens and 3 types of words per recording on average, though again speech was relatively denser in video vs. audio (1 vs. 3 hours depicted, respectively). Second, the number of talkers creeps up slightly with child age, with more talkers featured in audio recordings than video recordings. This is most readily explained by the greater portability of the audio recorders, and the daylong sampling they permitted. Finally, ~63% of nouns in the input are produced by mothers, ~13% by fathers, ~6% from toys and media, and the remainder by a mix of other talkers. This is particularly notable relative to the number of talkers featured in audio recordings; i.e, while there are often over 5 speakers, the majority of them provide a negligible amount of input relative to the child's parents.

In Figure 1, we compared the sheer volume of noun tokens in our dataset compared to CHILDES. Another way to compare the datasets is looking at unique noun types. As Figure 7 illustrates, SEEDLingS substantially extends the set of unique noun types when compared to CHILDES for the 6-17 month period. Of the 23,885 total unique noun types across both corpora, 8,184 appear only in CHILDES, 12,060 appear only in SEEDLingS, and 3,641 are shared between the two datasets. We note that for both corpora we count the nouns 'as spoken', but because of the different nature of the annotation processes this is not a perfect apples-to-apples comparison. E.g. CHILDES counts will include contractions (e.g. 'dog'll') while the SEEDLingS counts will include compounds and titles (e.g. "Go+Dogs+Go"); we encourage readers to keep this fuzziness in mind and see this as just an exploratory and imperfect comparison to give a sense of the data.

Discussion

Ways we have used the SEEDLingS-Nouns corpus

So far, the SEEDLingS-Nouns Dataset has let us tackle a wide variety of research questions, based either solely on the SEEDLingS-Nouns corpus data itself (e.g. Bergelson et al., 2018), or in conjunction with other measures (e.g. Bulgarelli & Bergelson, 2019). We mention a few findings below to give a sense of the methods and theoretical contributions of this corpus, and as a way to prime the pump for readers considering further use cases.

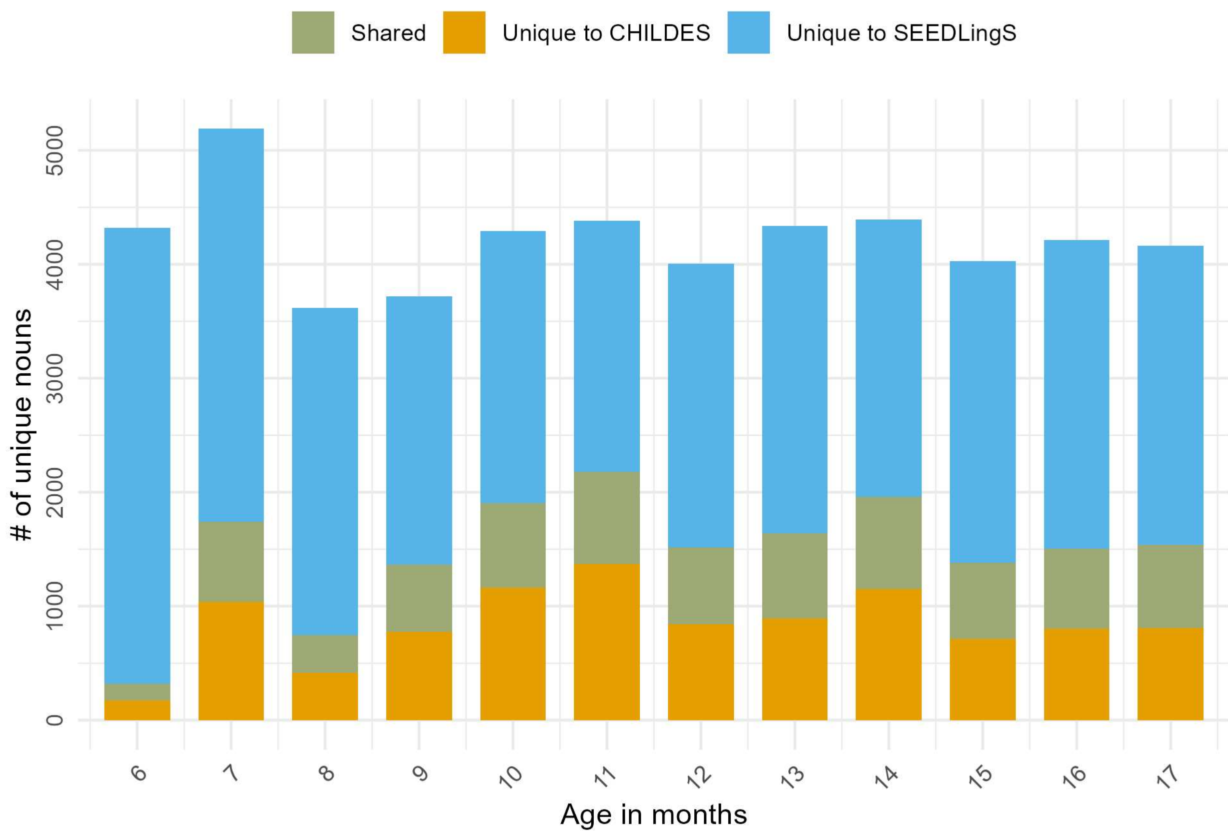

Figure 7. The number of unique noun types (distinct words) in the SEEDLingS-Nouns dataset and in the North American English portion of CHILDES, split by child age (6-17mo.) See text for details.

Work with this corpus has incorporated a range of methods including automated and manual acoustic analyses (Bulgarelli & Bergelson, 2019; Bulgarelli, Mielke, & Bergelson, 2022), integration across different facets of the observational data and its links to parent report (Moore et al., 2019), and computational approaches (Amatuni & Bergelson, 2017). The depth of the dataset has allowed us to analyze the effects of various demographic variables on word learning (e.g., gender in Dailey and Bergelson (2022), number of siblings in Laing and Bergelson (2024), and mothers' work schedules in Laing and Bergelson (2019)), though admittedly within a relatively homogeneous group of North American monolingual English-learning infants.

It has also lent itself to a number of analyses of the structure and features of infants' early language environments, and infants' learning and knowledge with them. For instance, the dataset has been used to analyze the source and effects of variability, across and within speaker and families. This includes morphological, phonetic, and acoustic measures of variability (Bulgarelli & Bergelson, 2024; Bulgarelli et al., 2021; Moore & Bergelson, 2024), which, alongside frequency, predict earlier word production. Computational work has so far analysed the broad timeline of noun learning in the dataset in relation to properties of the nouns heard in the input. For instance, in one line of work (Amatuni & Bergelson, 2017; Amatuni, He, & Bergelson, 2018) we found that the structural properties of the early noun input (including semantic and visual features) predict the timeline of acquisition of these nouns in the infants' vocabularies. The tagged noun data (e.g. object presence and utterance types) has also provided useful variables for the analysis of word learning and early input. For example, infants were better able to recognize nouns in the lab if they occurred with more frequent object presence in their own home-recorded data (Bergelson & Aslin, 2017), and infants with more siblings experienced less object presence in their input (Laing & Bergelson, 2024). Additionally, Egan-Dailey and Bergelson found that children who heard more nouns in imperative utterances (e.g., "drop that spoon!") during infancy scored lower on a standardized language skill assessment three years later (under review).

Finally, but not exhaustively, the dataset has been a useful testing ground for broader methodological questions. The combination of data sources (day-long audio, hour-long video, manual human annotations, LENA automated counts) has made it a valuable resource for validating approaches taken in the field more generally as represented by the three examples below. Bulgarelli and Bergelson (2019) provide further validation for LENA's automated speaker tags by showing moderate-to-strong agreement between automated and hand-coded tags. Bergelson et al. (2018) compare input measures in the day-long audio and hour-long video recordings to show that hour-long recordings capture a less typical, but more language-laden slice of infants' early language experience than recordings taken across an entire day.

More concretely, this research found that while audio recordings captured approximately 10 times more awake time than videos, the noun input in them was only 2-4 times greater. Thus, videos featured denser noun input, similar to ‘peak’ audio hours, though with relatively fewer declaratives and more questions. The hour-long videos thus provide a dense but somewhat different sample of infants’ language experiences rather than a “typical” one, highlighting the utility of considering different contexts.

Data use by others, limitations, and future directions

The work just mentioned largely focused on our own use of the SEEDLingS-Nouns data. However, the underlying recordings have also been used in a variety of other efforts, of which we highlight a few.

It has been used to establish best practices for data sharing, and manual transcription and annotation via the ACLEW Annotation Scheme (Soderstrom et al., 2021; VanDam et al., 2019) and for cross-corpus and cross-cultural comparisons of certain types of speech, such as rates of certain types of babbling, prevalence of various talkers, addressees, and objects (Bergelson et al., 2019, 2023; Bunce et al., 2024; Casey et al., 2022; Cychosz et al., 2021; Hitczenko et al., 2023). Cristia et al. (2024) assess the reliability of the proprietary LENA and open-source ACLEW pipelines across diverse datasets, showing low-to-moderate agreement in metrics and advising caution when using them to study individual variation. It’s further been used in a variety of speech-technology validations, evaluations, and advances (Cristia, Bulgarelli, & Bergelson, 2020; Cristia et al., 2021; Lavechin et al., 2022; Räsänen et al., 2019; Ryant et al., 2018; Schuller et al., 2019, 2017).Finally, the recordings in the corpus are also being used to drive theoretical advances in our understanding of what makes words more readily learnable at a mechanistic level (Beech, Bulgarelli, & Swingley, 2023), and how larger-scale forces like financial scarcity may connect with parent talk (Ellwood-Lowe, Foushee, & Srinivasan, 2022).

In the work just mentioned, as the originating lab for the dataset, we've played a range of roles from coauthors to data donors to interested observers. We welcome other roles, and in particular hope that the noun annotations in this dataset will continue to be used in many fruitful ways to test theory-driven questions regarding early word learning and its broader connections to other aspects of linguistic, cognitive, and social development. We hope too that they continue to permit and support validation of new speech technologies, an undervalued but critical component of new speech tool development.

While potential use cases for this data are many, we want to highlight that it does provide a very particular type of slice of everyday input. Firstly, all the usual caveats regarding generalization beyond this sample apply (i.e. other socio-linguistic and cultural contexts, other ages, typological factors, activity contexts etc. may render divergent results). But beyond this, we want to highlight that focusing on the set of concrete nouns we targeted in particular likely had consequences as well. A deeper dive into other kinds of nouns (e.g. social categories of people which we generally omitted), other lexical classes, particular activity contexts or syntactic constructions all might have shown different patterns in the input overall or over developmental time (within or beyond our 6-17mo focus). We would particularly welcome further cross-linguistic comparison with these caveats in mind, as concreteness, imageability, nominal biases, and caretaker-child interactions all provide a fertile ground for further comparative inquiry.

Conclusion

A decade of data collection, aggregation, cleaning, and curation has led to the SEEDLingS-Nouns dataset presented here. It makes a unique contribution in the depth with which it characterizes the dominant lexical class in the early vocabulary of monolingual English-learning infants: concrete nouns. Moreover, to help support the data's reuse, we provide many layers of metadata, supporting code and documentation, and details about the data's origins, processing, and uses to date. We look forward to seeing the contributions that ~360,000 nouns heard (or said) by the 44 SEEDLingS infants over the year we observed them, and the who what and when of their use, makes in advancing language science.

Acknowledgements

We'd particularly like to thank the undergraduate research assistants and research staff who helped collect, annotate, and clean this dataset across 3 institutions (Rochester, Duke, and Harvard) from 2014–2024. Space precludes listing each by name but please find a list of lab alumni (the majority of whom touched this dataset in some way!) here: https://bergelsonlab.com/people.html. We are grateful for your indefatigable help.

Declarations

Funding

The research was supported by the following grants: NIH grant 5DP5OD019812-05 to Elika Bergelson, NIH-NICHD grand F32 HD101216 to Federica Bulgarelli.

Conflicts of interest/Competing interests

None of the authors have any conflicts of interest to declare.

Ethics approval

The study (spanning data collection, annotation, and analysis) was approved by the Rochester, Duke, and Harvard IRBs, as relevant.

Consent to participate

Caregivers provided consent on behalf of their infants at an initial lab visit for the larger yearlong study through a process approved by the University of Rochester IRB.

Consent for publication

Caregivers additionally designated the level at which we may share their child's data and were given the option to change this designation at each visit. Parents also obtained signed permission from individuals who appeared in recordings in addition to parent and child.

Availability of data and materials

The dataset described here is available at doi.org/10.5281/zenodo.7709427 and github.com/bergelsonlab/seedlings-nouns

Due to privacy concerns inherent in the underlying naturalistic child-centered audio and video recordings, the raw data cannot be shared publicly. However, the majority of recordings are shared with authorized researchers, to the level of access that parents signed their explicit consent for. Videos are available through Databrary (Bergelson, 2016b) and audio-recordings through Homebank (Bergelson, 2017), two research data repositories that follow rigorous data safety and ethics standards.

These recordings only contain video and audio that has been appropriately screened for sensitive information. Any audio and/or video in the annotated time regions that contains direct identifiers or potentially private information has been silenced and/or covered. Additionally, any time regions that coders have not listened to have been removed, given the possibility of including the above types of information. Researchers interested in analyzing the recordings beyond our noun annotations are encouraged to the senior author (EB) with proposals; for our participants' privacy, unreviewed audio is not openly available, and the level of sharing authorized by participating families varies, as noted above.

Code availability

The article was written as an `RMarkdown` document using R package `papaja` (and many other packages). The document and supporting code and data are available on OSF (https://osf.io/r9pvn).

Authors' contributions

Initial project conception and funding (EB). Data collection (SK, ST, SE, EB). Data annotation, cleaning, checking, team management (LR, CM, AA, SK, CL, ST, FB, SE, HG, EB). Writing, maintaining, and updating code to work with and share the data (CM, AA, SK, SE, GB, SU, EK, EB). Creating project documentation (LR, CM, AA, SK, ST, SE, HG, EK, EB). Writing initial manuscript draft (EK, EB). Reviewing, revising, editing manuscript draft (all authors). N.B. Other than first and last authors, authors are listed in the order that they joined the project.

Open practices statement

All code and data needed to recreate this manuscript and all derived tables and figures is on OSF (https://osf.io/r9pvn). See the Declarations section for further details.

References

Allen, G. D. (1988). The PHONASCII system. *Journal of the International Phonetic Association*, *18*(1), 9–25.

Amatuni, A., & Bergelson, E. (2017). Semantic networks generated from early linguistic input. *bioRxiv*. https://doi.org/10.1101/157701

Amatuni, A., He, E., & Bergelson, E. (2018). *Preserved Structure Across Vector Space Representations*. arXiv. https://doi.org/10.48550/ARXIV.1802.00840

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking Quality and Quantity of Parental Linguistic Input to Child Language Skills: A Meta-Analysis. *Child Development*, *92*(2), 484–501. https://doi.org/10.1111/cdev.13508

Babineau, M., Carvalho, A. de, Trueswell, J., & Christophe, A. (2021). Familiar words can serve as a semantic seed for syntactic bootstrapping. *Developmental Science*, *24*(1), e13010. https://doi.org/10.1111/desc.13010

Barbaro, K. de, & Fausey, C. M. (2022). Ten Lessons About Infants' Everyday Experiences. *Current Directions in Psychological Science*, *31*(1), 28–33. https://doi.org/10.1177/09637214211059536

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123. https://doi.org/10.1017/S0305000900008680

Beech, C., Bulgarelli, F., & Swingley, D. (2023). Relating referential transparency and phonetic clarity in the SEEDLingS corpus [Public registration]. Retrieved from https://osf.io/7ydb3/resources

Benedict, H. (1979). Early lexical development: comprehension and production. *Journal of Child Language*, *6*(2), 183–200. https://doi.org/10.1017/S0305000900002245

Bergelson, E. (2016b). *SEEDLingS corpus*. Databrary.

Bergelson, E. (2016a). *SEEDLingS corpus*.

Bergelson, E. (2017). *Bergelson seedlings HomeBank corpus*. HomeBank. https://doi.org/10.21415/T5PK6D

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2018). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1). https://doi.org/10.1111/desc.12715

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. https://doi.org/10.1073/pnas.1712966114

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What Do North American Babies Hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724. https://doi.org/10.1111/desc.12724

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., … Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., … Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, French, hebrew, italian, korean, and american english. *Child Development*, *75*(4), 1115–1139. https://doi.org/10.1111/j.1467-8624.2004.00729.x

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and

Variability in Children's Word Learning Across Languages. *Open Mind*, *3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. https://doi.org/10.1016/s0010-0277(01)00122-6

Bulgarelli, F., & Bergelson, E. (2019). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavior Research Methods*, *52*(2), 641–653. https://doi.org/10.3758/s13428-019-01265-7

Bulgarelli, F., & Bergelson, E. (2024). Linking acoustic variability in the infants' input to their early word production. *Developmental Science*. https://doi.org/10.1111/desc.13545

Bulgarelli, F., Mielke, J., & Bergelson, E. (2021). Quantifying Talker Variability in North-American Infants' Daily Input. *Cognitive Science*, *46*(1). https://doi.org/10.1111/cogs.13075

Bulgarelli, F., Mielke, J., & Bergelson, E. (2022). Quantifying Talker Variability in North-American Infants' Daily Input. *Cognitive Science*, *46*(1), e13075. https://doi.org/10.1111/cogs.13075

Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Alam, F., … Casillas, M. (2024). A cross-linguistic examination of young children's everyday language experiences. *Journal of Child Language*, 1–29. https://doi.org/10.1017/S030500092400028X

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. https://doi.org/10.1073/pnas.1309518110

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

https://doi.org/10.1016/S0010-0277(96)00793-7

Casey, K., Elliott, M., Mickiewicz, E., Silva Mandujano, A., Shorter, K., Duquette, M., … Casillas, M. (2022). Sticks, leaves, buckets, and bowls: Distributional patterns of children's at-home object handling in two subsistence societies. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). Retrieved from https://escholarship.org/uc/item/6wx2x30s

Casey, K., Potter, C. E., Lew-Williams, C., & Wojcik, E. H. (2023). Moving beyond "nouns in the lab": Using naturalistic data to understand why infants' first words include uh-oh and hi. *Developmental Psychology*, *59*(11), 21622173. https://doi.org/10.1037/dev0001630

Casillas, M., Foushee, R., Méndez Girón, J., Polian, G., & Brown, P. (2024). Little evidence for a noun bias in Tseltal spontaneous speech. *First Language*, *44*(6), 600–628. https://doi.org/10.1177/01427237231216571

Choi, S. (2000). Caregiver input in english and korean: Use of nouns and verbs in book-reading and toy-play contexts. *Journal of Child Language*, *27*(1), 69–96. https://doi.org/10.1017/s0305000999004018

Coffey, J. R., Zeitlin, M., Crawford, J., & Snedeker, J. (2024). It's All in the Interaction: Early Acquired Words Are Both Frequent and Highly Imageable. *Open Mind*, *8*, 309–332. https://doi.org/10.1162/opmi_a_00130

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis System Segmentation and Metrics: A Systematic Review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Gautheron, L., Zhang, Z., Schuller, B., Scaff, C., Rowland, C., … Soderstrom, M. (2024). Establishing the reliability of metrics extracted from long-form recordings using LENA and the ACLEW pipeline. *Behavior Research Methods*. https://doi.org/10.3758/s13428-024-02493-2

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., … Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., … Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*, *24*(5). https://doi.org/10.1111/desc.13090

Dailey, S., & Bergelson, E. (2022). Talking to talkers: Infants' talk status, but not their gender, is related to language input. *Child Development*, *94*(2), 478–496. https://doi.org/10.1111/cdev.13872

Dailey, S., & Bergelson, E. (2023). Talking to talkers: Infants' talk status, but not their gender, is related to language input. *Child Development*, *94*(2), 478–496. https://doi.org/10.1111/cdev.13872

Datavyu Team. (2014). *Datavyu: A Video Coding Tool*. Databrary Project, New York University.

Egan-Dailey, S., & Bergelson, E. (under review). *Early child measures outpredict input measures of preschool language skills in u.s. English learners.*

Ellwood-Lowe, M. E., Foushee, R., & Srinivasan, M. (2022). What causes the word gap? Financial concerns may systematically suppress child-directed speech. *Developmental Science*, *25*(1), e13151. https://doi.org/10.1111/desc.13151

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i. https://doi.org/10.2307/1166093

Franco, F., Suttora, C., Spinelli, M., Kozar, I., & Fasolo, M. (2022). Singing to infants matters: Early singing interactions affect musical preferences and facilitate vocabulary building. *Journal of Child Language*, *49*(3), 552–577. https://doi.org/10.1017/S0305000921000167

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language*, *2*, 301–334.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing Children's Home Language Environments Using Automatic Speech Recognition Technology. *Communication Disorders Quarterly*, *32*(2), 83–92. https://doi.org/10.1177/1525740110367826

Hitczenko, K., Bergelson, E., Casillas, M., Colleran, H., Cychosz, M., Grosjean, P., … Cristia, A. (2023). The development of canonical proportion continues past toddlerhood. *Proceedings of the 20th International Conference of Phonetic Sciences.* Prague, Czech Republic.

Huttenlocher, J. (1974). *The origins of language comprehension 1.* Routledge.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' Memory for Spoken Words. *Science*, *277*(5334), 1984–1986. https://doi.org/10.1126/science.277.5334.1984

Laing, C., & Bergelson, E. (2019). Mothers' Work Status and 17-Month-Olds' Productive Vocabulary. *Infancy*, *24*(1), 101–109. https://doi.org/10.1111/infa.12265

Laing, C., & Bergelson, E. (2020). From babble to words: Infants' early productions match words and objects in their environment. *Cognitive Psychology*, *122*, 101308. https://doi.org/10.1016/j.cogpsych.2020.101308

Laing, C., & Bergelson, E. (2024). Analyzing the effect of sibling number on input and output in the first 18 months. *Infancy*, *29*(2), 175–195.

https://doi.org/10.1111/infa.12578

Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., … Bredin, H. (2022). *Brouhaha: Multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation*. arXiv. https://doi.org/10.48550/arXiv.2210.13248

Leech, K. A., McNally, S., Daly, M., & Corriveau, K. H. (2022). Unique effects of book-reading at 9-months on vocabulary development at 36-months: Insights from a nationally representative sample of Irish families. *Early Childhood Research Quarterly*, *58*, 242–253. https://doi.org/10.1016/j.ecresq.2021.09.009

Luo, R., Masek, L. R., Alper, R. M., & Hirsh-Pasek, K. (2022). Maternal question use and child language outcomes: The moderating role of children's vocabulary skills and socioeconomic status. *Early Childhood Research Quarterly*, *59*, 109–120. https://doi.org/10.1016/j.ecresq.2021.11.007

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Meylan, S. C., & Bergelson, E. (2021). Learning Through Processing: Toward an Integrated Approach to Early Word Learning. *Annual Review of Linguistics*, *8*(1), 77–99. https://doi.org/10.1146/annurev-linguistics-031220-011146

Miller, G. A. (1990). The Place of Language in a Scientific Psychology. *Psychological Science*, *1*(1), 7–14. https://doi.org/10.1111/j.1467-9280.1990.tb00059.x

Moore, C., & Bergelson, E. (2024). Wordform variability in infants' language environment and its effects on early word learning. *Cognition*, *245*, 105694. https://doi.org/10.1016/j.cognition.2023.105694

Moore, C., Dailey, S., Garrison, H., Amatuni, A., & Bergelson, E. (2019). Point, walk, talk: Links between three early milestones, from observation and parental report. *Developmental Psychology*, *55*(8), 1579–1593. https://doi.org/10.1037/dev0000738

Quam, C., Knight, S., & Gerken, L. (2017). The Distribution of Talker Variability Impacts Infants' Word Learning. *Laboratory Phonology*, *8*(1).

https://doi.org/10.5334/labphon.25

Räsänen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., … Soderstrom, M. (2019). Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech. *Speech Communication*, *113*, 63–80. https://doi.org/10.1016/j.specom.2019.08.005

Rebuschat, P., Meurers, D., & McEnery, T. (2017). Language Learning Research at the Intersection of Experimental, Computational, and Corpus-Based Approaches. *Language Learning*, *67*(S1), 6–13. https://doi.org/10.1111/lang.12243

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x.Speaker

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. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Rowland, C., Durrant, S., Peter, M., Bidgood, A., Pine, J., & Jago, L. S. (2018). *The Language 0-5 Project*. https://doi.org/10.17605/OSF.IO/KAU5F

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. https://doi.org/10.1073/pnas.1419773112

Ryant, N., Bergelson, E., Church, K., Cristia, A., Du, J., Ganapathy, S., … Yu, Z. (2018). Enhancement and Analysis of Conversational Speech: JSALT 2017. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5154–5158. https://doi.org/10.1109/ICASSP.2018.8462468

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science (New York, N.Y.)*, *274*, 1926–1928.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., … Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics

Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Interspeech 2019*, 2378–2382. ISCA. https://doi.org/10.21437/Interspeech.2019-1122

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., … Zafeiriou, S. (2017). The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring. *Interspeech 2017*, 3442–3446. ISCA. https://doi.org/10.21437/Interspeech.2017-43

Slone, L. K., Abney, D. H., Smith, L. B., & Yu, C. (2023). The temporal structure of parent talk to toddlers about objects. *Cognition*, *230*, 105266. https://doi.org/10.1016/j.cognition.2022.105266

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience. *Collabra: Psychology*, *7*(1), 23445. https://doi.org/10.1525/collabra.23445

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

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*, *20*(6), 10.1111/desc.12456. https://doi.org/10.1111/desc.12456

Tincoff, R., & Jusczyk, P. W. (1999). Some Beginnings of Word Comprehension in 6-Month-Olds. *Psychological Science*, *10*(2), 172–175. https://doi.org/10.1111/1467-9280.00127

VanDam, M., De Palma, P., Soderstrom, M., Casillas, M., Cristia, A., Bergelson, E., … MacWhinney, B. (2019). Daylong acoustic recordings of family and child speech using the HomeBank database. *The Journal of the Acoustical Society of America*, *145*(3), 1729–1729. https://doi.org/10.1121/1.5101352

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D.,

& MacWhinney, B. (2016). HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. *Seminars in Speech and Language*, *37*(02), 128–142. https://doi.org/10.1055/s-0036-1580745

Vihman, M. M., & McCune, L. (1994). When is a word a word? *Journal of Child Language*, *21*(3), 517–542. https://doi.org/10.1017/s0305000900009442

Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., & Song, H. (2013). Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child Development Perspectives*, *7*(3), 155–159.

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *WIREs Cognitive Science*, *13*(4). https://doi.org/10.1002/wcs.1596