An automated classifier for periods of sleep and target-child-directed speech from LENA recordings

Janet Y. Bang
San José State University, USA

George Kachergis
Stanford University, USA

Adriana Weisleder
Northwestern University, USA

Virginia A. Marchman
Stanford University, USA

Abstract: Some theories of language development propose that children learn more effectively when exposed to speech that is directed to them (target child directed speech, tCDS) than when exposed to speech that is directed to others (other-directed speech, ODS). During naturalistic daylong recordings, it is useful to identify periods of tCDS and ODS, as well as periods when the child is awake and able to make use of that speech. To do so, researchers typically rely on the laborious work of human listeners who consider numerous features when making judgments. In this paper, we detail our efforts to automate these processes. We analyzed over 1,000 hours of audio from daylong recordings of 153 English- and Spanish-speaking families in the U.S. with 17- to 28-month-old children that had been previously coded by human listeners for periods of sleep, tCDS, and ODS. We first explored patterns of features that characterized periods of sleep, tCDS, and ODS. Then, we evaluated two classifiers that were trained using automated measures generated from LENA™, including frequency (AWC, CTC, CVC) and duration (meaningful speech, distant speech, TV, noise, silence) measures. Results revealed high sensitivity and specificity in our sleep classifier, and moderate sensitivity and specificity in our tCDS/ODS classifier. Moreover, model-derived predictions replicated previously-published findings showing significant and positive links between tCDS, but not ODS, and children's later vocabularies (Weisleder & Fernald, 2013). This work offers promising tools for streamlining work with daylong recordings, facilitating research that aims to better understand how children learn from everyday speech environments.

Keywords: child-directed speech, other-directed speech, LENA, daylong recordings, automated classifier

Corresponding author(s): Janet Bang, Department of Child and Adolescent Development, San José State University, One Washington Square, San José, CA, 95192, USA. Email: janet.bang@sjsu.edu

ORCID ID(s): Janet Y. Bang: https://orcid.org/0000-0002-6014-3009
George Kachergis: https://orcid.org/0000-0003-4153-4167
Adriana Weisleder: https://orcid.org/0000-0001-6094-8424
Virginia A. Marchman: https://orcid.org/0000-0001-7183-6743

Citation: Bang, J.Y., Kachergis, G., Weisleder, A., & Marchman, V. A. (2023). An automated classifier for periods of sleep and target-child-directed speech from LENA recordings. *Language Development Research, 3*(1), 211–248. https://doi.org/10.34842/xmrq-er43

Introduction

Speech environments vary across children in numerous ways. The ability to document variation in children's naturally-occurring speech environments has been greatly assisted by technology that can capture, store, and process large amounts of audio data (e.g., an entire day). One notable example is the LENA digital language processor and software system (Gilkerson et al., 2017; Gilkerson & Richards, 2020). The recorder is worn inside a child's front shirt pocket and records the audio environment around the child, with each recording storing up to 16 hours of audio. The LENA software applies machine-learning algorithms to identify speech from children and adults that is "meaningful" or "near and clear" to the child (Cristia et al., 2021; Gilkerson & Richards, 2020). Summary reports provide estimates of the number of adult words (Adult Word Count, AWC), child vocalizations (Child Vocalization Count, CVC), and conversational turns (Conversational Turn Count, CTC), as well as the duration of time with meaningful speech, distant speech, TV/electronic media, non-speech noise (e.g., fan), and silence. A number of studies in different languages have compared these estimates to counts derived from human transcription and have reported mixed findings for the validity of LENA measures, with AWC, CTC, CVC among the most widely studied (Busch et al., 2018; Canault et al., 2016; Ferjan Ramirez et al., 2023; Gilkerson et al., 2015; Lehet et al., 2021; Soderstrom & Wittebolle, 2013; VanDam & Silbert, 2016; for a systematic review and meta-analyses of validation studies see Cristia et al., 2020).

Studies with LENA have been conducted in numerous languages and sociocultural settings. Most of these studies use LENA's estimates of AWC, CTC, and CVC to investigate how young children's language environments might support their language development, particularly by examining the amount and types of speech that are available to the child. Although the automated speech counts provided by LENA are useful, they are not sufficient to characterize many aspects of children's speech environments that are thought to be relevant for language learning. For example, segments with relatively high AWC values may indicate interactions when an adult is engaging verbally with their child (i.e., target-child-directed speech, tCDS). But these segments could also reflect periods in which multiple adults are talking to each other near the child, without any of the adults speaking directly to the child (i.e., other-directed speech, ODS). Similarly, some portions of the day may be characterized by high values for silence. These long periods of silence could reflect times when no speech is addressed to the child even though the child is awake and available to experience that speech (e.g., the caregiver is not interacting with the child or they are engaging nonverbally). Or, these periods could reflect times when the child is sleeping and no adults are present. These different scenarios have been proposed to play different roles in language learning and are of theoretical interest to many researchers. Yet, the LENA algorithms/measures do not currently distinguish between them.

Deriving estimates of the child-directed vs. other-directed nature of the speech that children hear is particularly important for our understanding of how children learn language from their speech environment (Dailey & Bergelson, 2022). A growing body

of work has proposed that target-child-directed speech, more so than other-directed speech, supports language development (Goldstein & Schwade, 2008; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). Relatedly, when caregivers engage verbally with young children, the extent to which they use a child-directed register, i.e., speech characterized by certain acoustic, prosodic, lexical, and morphosyntactic properties, has been proposed to be particularly conducive for learning (Fernald et al., 1989; Quigley et al., 2019; Singh et al., 2009; Soderstrom, 2007; Stärk et al., 2022). These studies exemplify the rapidly growing interest in identifying and characterizing periods of target-child-directed speech within daylong recordings.

Child-directed Versus Other-directed Speech

The construct of child-directed speech is central to theories that aim to explain how children learn language from social interactions (Csibra & Gergely, 2009; Tomasello, 1995). However, communities vary widely in how much speech is directed to children and how much speech is spoken around the child but not directed to them (Casillas et al., 2019; Ochs & Schieffelin, 1984; Shneidman & Goldin-Meadow, 2012). Despite this variability, cross-cultural work finds that key language milestones (e.g., onset of first words and multi-word utterances) emerge around the same age in a variety of communities (Casillas et al., 2019; Crago et al., 1997). Such findings raise questions regarding whether any speech in children's environments, whether it is addressed to them or not, can support their language acquisition.

Indeed, lab-based experimental studies have demonstrated that children can learn new words from speech that is not explicitly directed to them. For example, Akhtar and colleagues (2001) found that 1- to 2-year-old children were able to learn novel nouns and verbs when observing two adults play a game. Other studies varied the degree of joint attention between speaker and learner, such as having speakers turn their backs to infants during a word learning episode, replicating the finding that children can learn new words even in such contexts (Gampe et al., 2012). In contrast, some research examining speech in natural environments reports that target-child-directed speech, more so than other-directed speech, is associated with children's vocabulary development (Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). For example, using LENA recordings with 29 Spanish-speaking families in the U.S., Weisleder and Fernald (2013) coded for periods of target-child-directed speech, i.e., speech directed to the target child in one-on-one interactions or with others, versus overheard speech[1], i.e., speech directed to other adults or children other than the target child. Using AWCs from LENA when the child was 19 months, they found that the number of adult words in periods with target-child-directed speech was related to children's vocabulary size at 25 months, while the number of adult words in periods

[1] Weisleder & Fernald (2013) and Shneidman & Goldin-Meadow (2012) used the term "overheard" speech. We use 'other-directed speech' as a more conservative term (Casillas et al., 2019), since it is unclear whether children do or do not hear speech when it is directed to others.

with other-directed speech was not. Similar findings were observed in Shneidman and Goldin-Meadow (2012), where the amount of target-child-directed, but not overheard, speech was associated with child vocabulary in Yucatec-Mayan-speaking families in subsistence farming communities in Mexico. Collectively, these studies reveal mixed findings about the differential roles of target-child-directed and other-directed speech in young children's language learning.

When caregivers engage with young children, they sometimes change their speech register, producing a type of speech colloquially referred to as "baby talk", "parentese", and which researchers refer to as "infant-directed speech (IDS)." Numerous acoustic, prosodic, phonological, lexical, grammatical, and pragmatic features have been noted to differentiate this child-directed register from adult-directed registers (Hilton et al., 2020; Soderstrom, 2007). Moreover, speech that is characterized by features of IDS has been suggested to be especially supportive of children's speech and language acquisition (Byers-Heinlein et al., 2021; Fernald et al., 1989; Singh et al., 2009; Snow, 1977). For example, a recent multi-continent collaboration demonstrated that speech characterized by the acoustic and phonological features of North American English IDS was preferred over speech spoken in an adult-directed register by both mono- and bilingually-exposed infants (Byers-Heinlein et al., 2021). These results were interpreted to suggest that acoustic features associated with IDS may be more effective at attracting infants' attention and thereby, can better support learning, particularly when young children are developing their early language skills. However, there is continued debate about the relative role of child- and adult-directed speech registers in children's language learning across linguistic and cultural contexts (Solomon, 2011; Cox et al., 2022).

LENA's View of the Auditory Environment

The main goal of the LENA system is to identify vocalizations from the child wearing the recorder and nearby adults, while excluding all other sounds (Gilkerson & Richards, 2020). The software uses various acoustic features to segment the audio recording and label the sounds into one of eight main categories: key child (the child wearing the recorder), adult female, adult male, electronic media (e.g., TV), other child, distant or overlapping speech, noise, and silence. The result of this process is an "Interpreted Time Segments" (ITS) file, which is in essence a diarization file (Xu et al., 2009). The ITS file is written in standard XML format and can be exported from the LENA software for each recording. The ITS file contains all the segmentation/diarization information, including the duration of each sound and its intensity (loudness).

In addition to segmenting and labeling the audio, LENA also estimates the frequency of adult words (AWC), adult-child conversational turns (CTC) and child vocalizations (CVC). To do this, LENA does not attempt to recognize actual words; instead, the algorithm estimates the number of words based on information in the speech signal, such as segment duration, syllable count, and consonant distribution (Gilkerson & Richards, 2020). These word and vocalization frequencies are estimated only from LENA's three primary speaker labels (adult female, adult male, and key child), or

what LENA calls "meaningful speech." No word/vocalization counts are estimated for other children or for distant/overlapping speech. All vocalization counts include speech-like vocalizations separated by a 300 ms break, but exclude respiratory (e.g., breathing) and digestive sounds (e.g., burping). These frequencies are exported as part of the ITS file. In addition, users can export summary-level reports from LENA, which provide word and vocalization counts (AWC, CTC, CVC) over a particular unit of time (e.g., 5 minutes, or 1 hour), as well as time-based measures of the amount of time (minutes) within that unit that contain meaningful speech (i.e., speech that is 'near and clear'), distant/overlapping speech, TV/electronic media, non-speech noise (e.g., fan), and silence. These summary reports are used by most LENA users to characterize the child's speech environment, as they provide useful information about the amount of adult speech the child hears throughout the day, the child's own vocalizations, and the number of conversational exchanges between the child and adult(s). The AWC measure is the most-widely used, as well as the most reliable/accurate of these measures. However, this measure does not distinguish whether the adult speech is directed to the child or just spoken in the child's vicinity. Additionally, LENA does not identify whether the speech is characterized by prosodic and acoustic features of child-directed register, e.g., exaggerated intonation. Thus, to date, researchers interested in these distinctions have had to rely on manual annotation.

Identifying periods of sleep, target- and other-directed speech in daylong recordings

Manual annotation of LENA recordings requires that human listeners identify periods of sleep, target-child-directed, and other-directed speech by attending to numerous cues that are available on the audio recording (Weisleder & Fernald, 2013). However, these efforts are highly labor and time intensive. Though there are emerging tools to support the rigor and efficiency of this type of manual coding (Cychosz et al., 2021; Mendoza & Fausey, 2021), efforts to automate steps in this process are also in critical need. Additionally, in some cases, ethical considerations prevent researchers from listening to the recordings (Cychosz et al., 2020).

Recent work has demonstrated progress in automating speech classifications as infant/child- vs. adult-directed registers from daylong recordings (De Palma & VanDam, 2017; Schuller et al., 2017) or laboratory stimuli (Räsänen et al., 2018; Schuster et al., 2014), mainly by focusing on the acoustic and phonetic features of speech. However, no studies to our knowledge have demonstrated the extent to which we can reliably classify whether speech was directed to the target child or not from daylong recordings, regardless of register. Thus, tools that enable classification of periods of target-child-directed and other-directed speech from features that are automatically extracted from the recordings could expand the range of cases in which such features can be examined.

Figure 1 depicts examples from three children's daylong recordings (from Weisleder & Fernald, 2013), illustrating the automated AWC estimates (adult females and adult males) per 5-min audio segment across the day. Not surprisingly, the AWC values in

each segment for a given child vary considerably across the day, and the mean AWC values that are averaged across the day also vary across the three children. To determine which AWC values reflect tCDS rather than other-directed speech, human listeners judged each 5-minute segment first as whether the child was sleeping and, if not, whether the adult speech during the segment was more than 50% tCDS or ODS. Notably, removing ODS segments changed the estimates of overall speech to the child across the day substantially for some children, but less so for others.

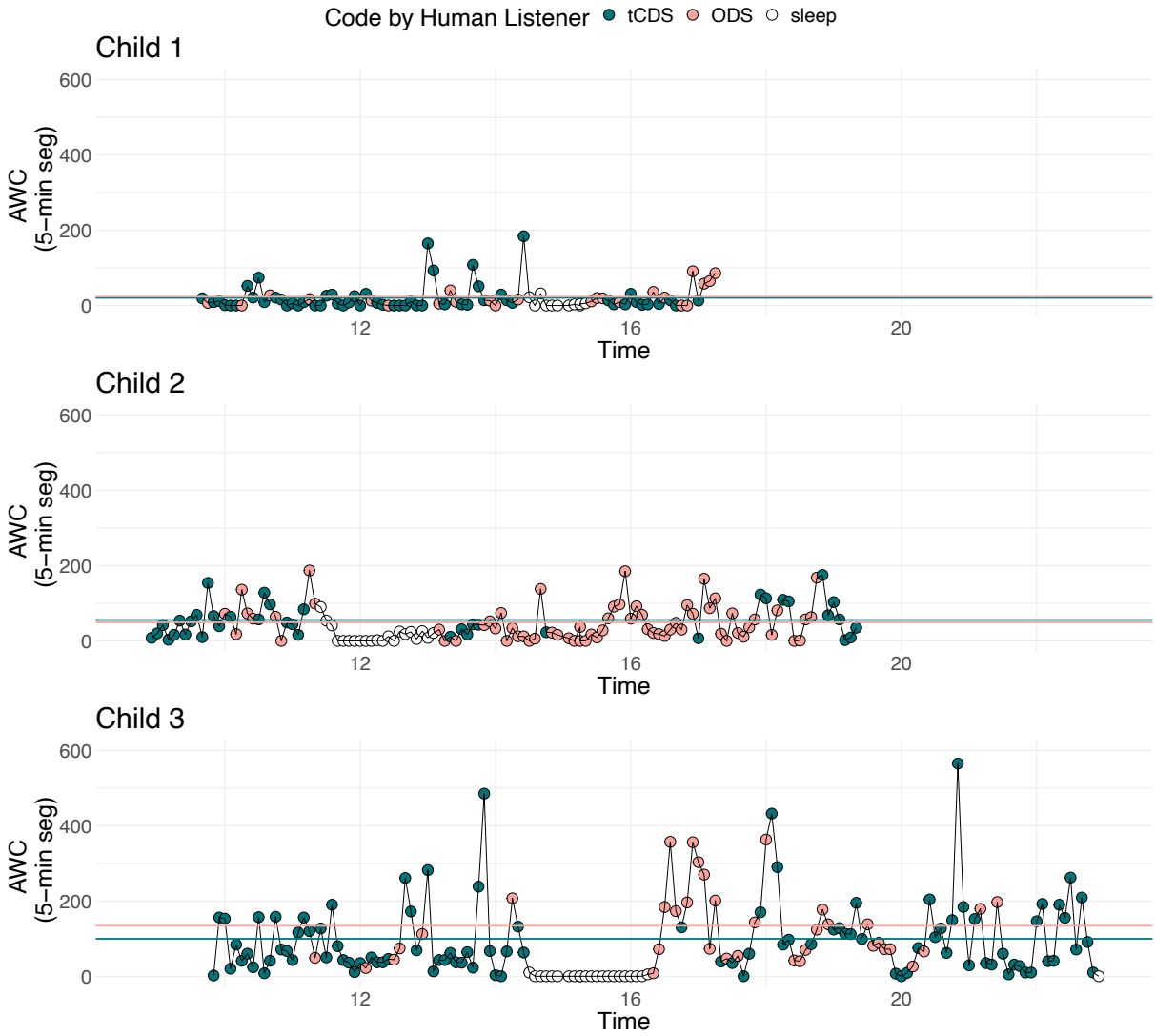


Figure 1. ***Example profiles of three children's AWC counts per 5-minute segment across their daylong recordings. Note: Green dots represent segments judged by human listeners to be more than 50% target-child-directed speech; Light pink dots represent segments judged to be more than 50% other-directed speech; White dots represent segments when the child was sleeping as judged by human listeners. Horizontal lines depict the average tCDS (green) or ODS (pink) counts computed over the entire recording.***

As this figure shows, periods of tCDS or ODS were not easily differentiated by AWC (i.e., segments that were identified as tCDS and ODS had a range of both high and low AWC values). Thus, to differentiate periods of tCDS and ODS, it may be more productive to examine multiple measures in combination. For example, a given 5-minute segment may be more likely to be tCDS when that segment's AWC value is interpreted in conjunction with relatively high values of CTC or CVC. Similarly, one could predict that periods of children sleeping would be characterized by both low values of AWC and low values of CVC or CTC. By combining across measures, we can gain insights into which features conspire to reflect periods of sleep or of tCDS versus ODS and how best to identify them automatically.

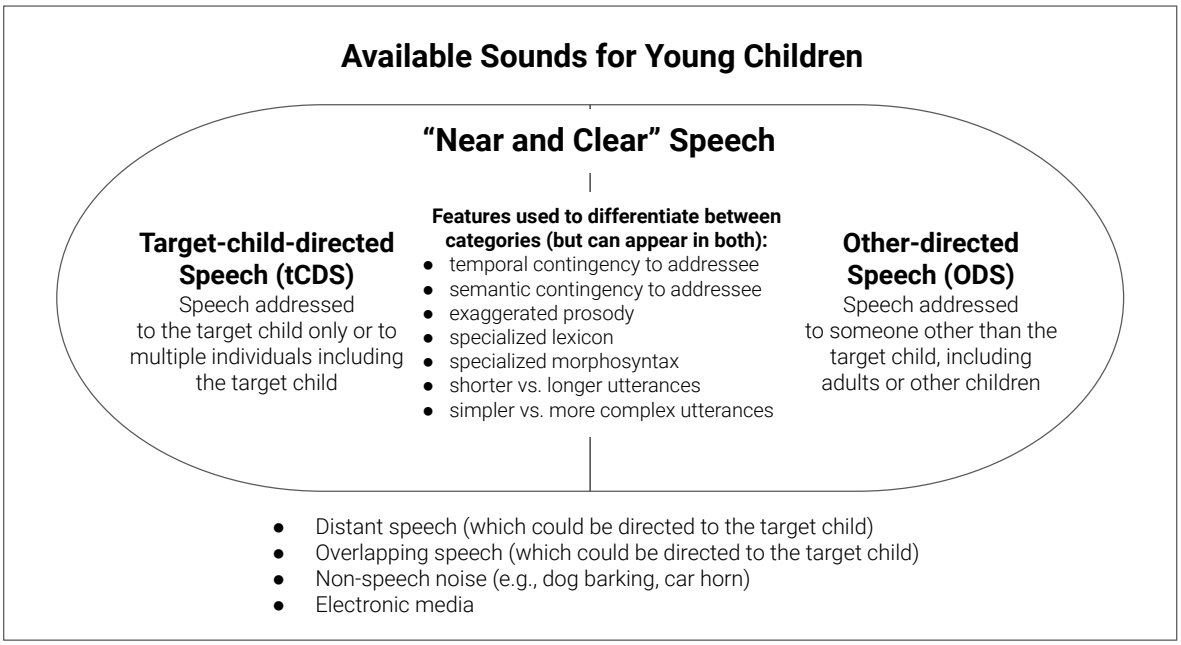
Current Study

This study examined ways to facilitate the identification of periods of target-child-directed vs. other-directed speech in daylong LENA recordings, as well as periods when children are awake versus sleeping, using only the automated measures provided in the standard LENA summary reports. By focusing on the automatically-generated LENA measures, we seek to develop classification tools that require minimal additional processing of the data and that can be easily integrated into a workflow.

Figure 2 provides our conceptualization of target-child-directed vs. other-directed speech. Looking only at the speech that is "near and clear," i.e., potentially audible by the target child, we defined tCDS as all speech that is directed to the target child, either individually or part of a group. In contrast, ODS is defined as all "near and clear" speech that is addressed to others. Note that other features cross-cut these categories. For example, while tCDS is more likely to be characterized by short utterances and child–directed prosody, there are times when speech that is clearly directed to a child does not fit that characterization. Analogously, while ODS may be more likely to be spoken in adult-directed prosody, there are also times and contexts when ODS might share many of the features, e.g. exaggerated prosody, characteristic of child-directed speech. Our goal was to develop a tool that could effectively identify periods of all speech directed to the target child, some of which may use a CDS register and some which may not. By focusing more generally on the function, rather than the features of speech, we align our research questions with the more general theoretical proposal that children learn language through interactions with others, and that they may learn best from language input that is contingent on or relevant to their vocalizations, actions, and/or attentional focus (Goldstein & Schwade, 2008; McGillion et al., 2013; Tamis-LeMonda et al., 2014; Tomasello, 1995; Yurovsky, 2018).

Our approach is as follows. We first conducted preliminary analyses to explore how the core frequency count measures, i.e., AWC, CTC, and CVC, worked in combination to predict periods of target-child-directed vs. other-directed speech. Using data from recordings of 29 Spanish-speaking families in the U.S. (from Weisleder & Fernald, 2013), we conducted logistic regressions to assess the degree to which variation in these measures was associated with whether a particular 5-minute segment was classified as tCDS or ODS by human coders. Next, we compiled data across several studies

of English- and Spanish-speaking families in the U.S. (n = 153), applying more complex machine-learning classifiers that combined the frequency (AWC, CTC, CVC) and time-based measures (meaningful speech, distant speech, TV, noise, and silence) to



identify periods of sleep, tCDS, and ODS that had been previously identified by human coders. We first used cluster analyses to examine how these multiple LENA features hung together and then developed two classifiers, one for distinguishing periods when the target child was asleep versus awake and another for distinguishing periods of primarily tCDS versus ODS.

Figure 2. ***Conceptualization of tCDS and ODS in our study. Note: "Near and clear" describes the audible speech from the perspective of the child wearing the recorder (Cristia et al., 2021; Gilkerson & Richards, 2020), which we define as the target child.***

Performance of both the sleep and tCDS/ODS classifier were evaluated based on the concordance with the human coders, defined in terms of both the sensitivity and specificity of the model predictions in comparison to human coders (ground truth). These estimates provide a standard measure of classification ability reflecting the degree to which the classifiers can distinguish both negative and positive values of each category. For the tCDS/ODS classifier, we also evaluated its performance in terms of its ability to replicate previously published links between variation in adult word counts and children's later vocabulary outcomes. In particular, Weisleder & Fernald (2013) reported stronger correlations between parent-reported vocabulary size and AWC values derived from 5-minute segments categorized as primarily tCDS, compared to those based on 5-minute segments identified as being primarily ODS. If a similar pattern of correlations is found with classifier-based values, this would provide some assurance that the classifier is capturing dimensions of children's language

input that are analogous to those identified by human coders.

Methods

Participants

Participants were families and their 17- to 28-month-old children from 79 English- and 74 Spanish-speaking households in the U.S. In total, families contributed over 1,000 recorded hours of LENA recordings (12,936 5-min segments). Descriptives are shown in Table 1. Data analyzed were collected between 2008-2015. Recruitment information is reported elsewhere (Fernald et al., 2013; Marchman et al., 2020; Weisleder & Fernald, 2013).

Table 1. ***Descriptive statistics of participants and recordings in the five different samples.***

Sample	n	Lang.	Age range (mo)	Mat. Ed range (y)	Total recording length in hours Mean (SD)	Seg. dur (min)	Num seg. incl.
1	27	En	18 - 19	12 - 18	10.62 (2.29)	5	3491
2	29*	En	17 - 19	10 - 18	9.32 (2.52)	5	3275
3	45	En	23 - 26	10 - 18	11.05 (3.22)	5**	1891
4	29	Sp	18 - 20	4 - 16	10.67 (3.13)	5	2758
5	45	Sp	25 - 28	6 - 18	13.44 (3.68)	5**	1521

Note: *n = 22 from Sample 2 are also included in Sample 3 at a second time point, thus the total sample results in 153 unique families; En = English, Sp = Spanish. **10-minute segments rated by human coders were split into 5-minute segments for the purpose of our analyses.

Data collection and Coding

Across all studies, research staff obtained informed consent from caregivers and provided instructions of how to use LENA. Caregivers were asked to record on a "typical day." To respect families' privacy, caregivers were told that they could pause the recording at their convenience. Recording instructions varied slightly across samples, but in all cases, families were given a single LENA recorder to use on a single day or across multiple days. All families were encouraged to record during all parts of the day. All recordings were cleaned following a standard lab protocol to exclude portions

of the recording when the LENA was not being used as recommended (e.g., the child was not wearing the vest, or the caregiver asked us not to listen to a period of the day.) Details about cleaned versus uncleaned recordings can be seen in Bang et al. (2022) and Weisleder & Fernald (2013).

Next, native speakers of each language coded segments of the audio-recording. For all samples, coders classified each segment as tCDS or ODS based on the most prevalent type of speech in that segment. For samples 1, 2, and 4 (see Table 1), human listeners listened to the entire recording and coded each 5 min segment as consisting of: sleep, primarily tCDS, primarily ODS, or a 50/50 split between tCDS or ODS. Segments of sleep were confirmed by environmental sounds (e.g., deep breathing). Segments identified as tCDS were those in which the majority (> 50%) of the surrounding adult speech (i.e., represented by the AWC value) was directed to the target child wearing the recorder, either addressed exclusively to the target child or inclusive of the target child (e.g., a speaker addressed a group that included the target child). Coders based their judgments on numerous features including the content of the speech, as well as exaggerated prosody, slower speech tempo, affect, perceived distance of the speaker relative to the child, environmental sounds, who responded to the speaker, and the activity of the interaction. Segments identified as ODS were those in which the majority of the speech was not directed to the target child nor inclusive of the target child. Split segments were those judged to have equal amounts of tCDS and ODS. For all preliminary analyses and the classifiers, we treated all 'split' segments as ODS, so that all segments coded as tCDS reflected segments with more than 50% target-child-directed speech.

For samples 3 and 5, a slightly revised protocol was followed. Here, coders first listened to potential periods of sleep based on information in a log book, targeting segments with consecutive low AWC values (AWC values = 0 for a minimum of 2 consecutive segments). If the child was confirmed to be sleeping, coders continued listening to segments prior to and after these segments to determine the beginning and end of periods of sleep. Next, families' highest AWC values were sorted in descending order based on 10-min segments, and coders rated each segment as primarily tCDS or ODS if approximately 70% of speech was either tCDS or ODS until six segments of primarily tCDS were identified per family (Bang et al., 2022). These 10-min segments were split into 5-min segments for the purpose of the current analyses, attributing the assigned code to each of the 5-min segments.

Reliability

To assess reliability of human coding, we determined the degree to which judgments of tCDS or ODS were consistent between two human raters. For each sample, we randomly selected 5 families (approximately 10 - 20% per sample, depending on the sample size) to be double-coded. For samples 1, 2, and 4, each family's recording was split into thirds and we randomly sampled five continuous 5-min segments from each block. Continuous segments were selected for double-coding in order to create coding

conditions that were analogous to those of the original coders who listened to the entire recording. For samples 3 and 5, we randomly selected ~eight 10-min segments for two families per sample, splitting the 10-min segments into 5-min segments, for a total of ~16 5-min segments coded by second raters. For purposes of reliability calculations, we excluded segments identified as splits during initial coding (n = 37, 9.7% for samples 1, 2, and 4) and other segments that were previously removed from analysis (n = 15, 3.9%). Judgments were compared between two raters (with different combinations of first and second raters), using K = 2 codes (tCDS or ODS), and raters coded independently (i.e., second raters had no knowledge of codes by first raters). Our coding protocol can be seen here: https://osf.io/qcj6r/.

For all samples, first raters were treated as the gold standard. Human raters judged each 5-min segment as having (a) no caregiver speech, (b) <less than 50% tCDS, (c) 50 - 70% tCDS, or (d) >70% tCDS. Segments rated as (a) or (b) were considered ODS; segments rated as (c) or (d) were considered tCDS. We evaluated our interrater reliability using Cohen's kappa and estimated rater accuracy (Bakeman, 2022). The value of "estimated rater accuracy" is determined from a simulation using the KappaAcc program (Bakeman, 2022), and reflects how accurate simulated observers would need to be to obtain the same observed kappa given the specifics of the data. Estimated rater accuracy provides a metric to judge "accurate enough" standards given the conditions of a dataset (e.g., number of raters and codes, frequency of different codes), rather than categorical cutoff points for Cohen's kappa. Table 2 reports that our Cohen's kappa for the total sample was .54 (80% agreement, uncorrected for chance). To produce a kappa of this value, the estimated rater accuracy suggests that simulated observers under similar circumstances (2 codes, 2 raters) would need to be 87% accurate (range across five samples = 77 - 90%).

Table 2. ***Reliability between first raters and second raters per sample and in total***

Sample	Language	n	Percent agreement (uncorrected)	Estimated rater accuracy	Cohen's kappa
1	En	5	85%	87%	.38
2	En	5	80%	89%	.61
3	En	5	79%	88%	.58
4	Sp	5	83%	90%	.65
5	Sp	5	73%	77%	.24
Total	En and Sp	25	80%	87%	.54

Results

Preliminary Analyses

Figure 3 illustrates the distributions of raw AWC, CTC, and CVC values per 5-min segment for tCDS or ODS segments using only data from Sample 4 (Weisleder & Fernald, 2013). To examine the degree to which the frequency measures of AWC, CTC, and CVC predicted the human-coded classifications of tCDS or ODS, we conducted hierarchical mixed effects logistic regression models. Models included a random intercept per participant and importantly, all frequency measures were converted to rates per minute and mean-centered within each family to allow interpretation of values as relative to each family's mean rates.

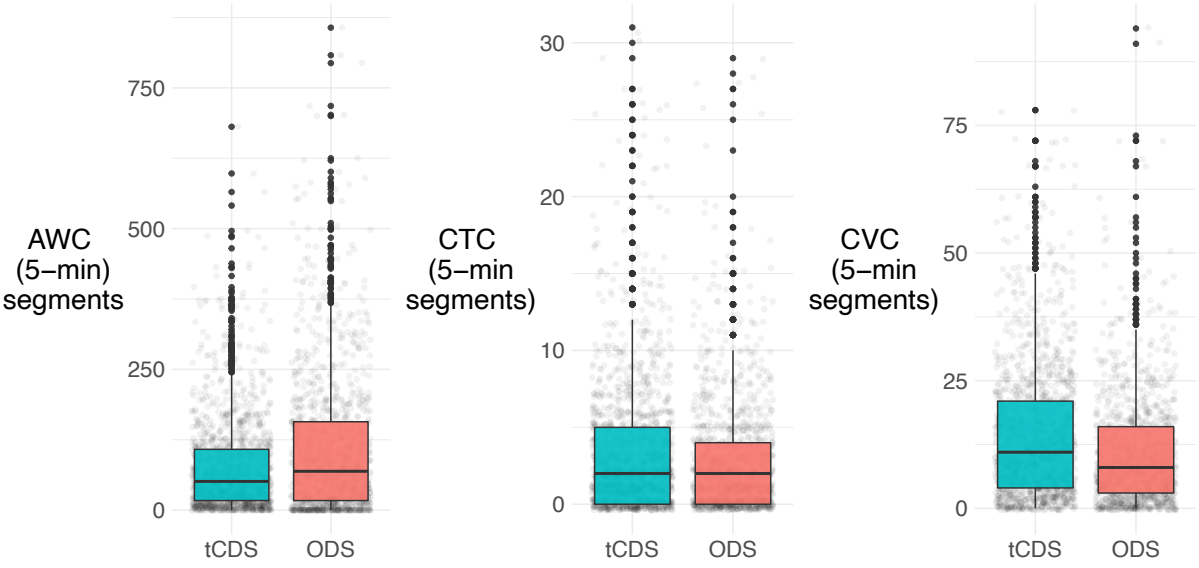


Figure 3. ***Boxplots of raw AWC, CTC, and CVC values by 5-min segments human-coded as tCDS or ODS using data from Sample 4. Note: Split segments were treated as ODS; segments identified as sleep were excluded.***

We found that each frequency measure, AWC/min, CTC/min, and CVC/min, independently contributed to the probability of a segment being classified as tCDS versus ODS. As seen in Figure 4, lower AWC rates (B = -.59, 95% CI = [-.72, -.46]) were associated with a higher probability of tCDS, indicating that segments that have higher than average AWC for a given family have a lower probability of being coded as tCDS by human coders. In contrast, higher rates of CTC (B = .36, 95% CI = [.21, .52]) and CVC (B = .39, 95% CI = [.26, .52]) were associated with a higher probability of tCDS, such that segments that were higher than average in CTC or CVC for a given family were more likely to be coded as tCDS. These findings indicate that each of the LENA frequency measures predicted the probability of tCDS, but did so in different directions. Moreover, because these measures are interrelated, it is likely that relations were more complex than these techniques could capture. Thus, we next recruited machine

learning techniques to explore the extent to which multiple LENA features, including both frequency- (AWC, CTC, and CVC) and time-based (e.g., minutes in meaningful speech, noise), could be used jointly to classify periods of sleep, tCDS or ODS.

Cluster Analyses

We next examined whether segments could be meaningfully clustered, which might suggest that a classifier based on thresholding multiple feature values (e.g., a decision tree) might work better than techniques that looked at predictors individually. We include speech frequency measures (AWC, CTC, and CVC) and time-based measures provided by LENA summary outputs (minutes in meaningful speech, distant speech, TV, noise, and silence), and examined how these measures clustered to predict human coding of the 5-min segment as periods of sleep, >50% tCDS, or > 50% ODS. Using an unsupervised clustering algorithm (k-means), we clustered all 12,936 segments according to their raw LENA values, considering k = {2,..,15} clusters. Table 3 shows the selected k = 7 clusters along with the proportion of each type of segment in the cluster and the mean values of LENA features for segments in that cluster. As shown in bold, Clusters 4 and 5 capture mostly sleep (64% and 53%) with low AWC, CTC, and CVC, but both clusters also include a moderate number of tCDS segments (22% and 30%). Note that Cluster 5 is also associated with high levels of noise (italicized), whereas Cluster 4 is associated with high levels of silence. The next two clusters in bold, Clusters 6 and 1, are both predominantly tCDS (73% and 60%) and cover 36.4% of the dataset, however, one has somewhat higher mean AWC, CTC, and CVC values than the other. Note also that these two clusters also contain many ODS segments. Next, we can note that Clusters 7 and 2 are comprised mostly of ODS segments. While both clusters are associated with low values of CTC and CVC, Cluster 7 is associated with high values of AWC, while Cluster 2 is not. Finally, Cluster 3, which looks much like the sleep clusters (4 and 5) in terms of low AWC, CTC, and CVC, is also associated with a higher level of TV than other clusters.

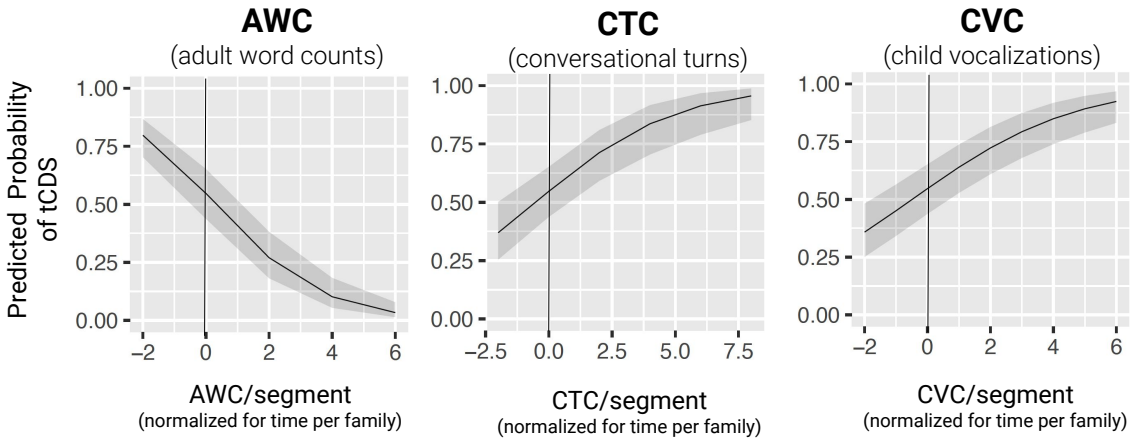


Figure 4. ***Predicted probabilities and confidence intervals (shaded region) of tCDS from AWC, CTC, and CVC, when holding each other measure at families' mean value (vertical line at 0).***

Overall, these cluster analyses showed that: 1) multiple LENA features captured meaningful variation between the clusters, as some features clustered together to correspond primarily to sleep, tCDS, or ODS, and yet 2) the clusters have significant overlap in tCDS and ODS, and to a lesser extent, sleep.

Table 3. ***Means of LENA features by cluster, annotated with proportion of sleep, tCDS, and ODS segments.***

		Category			LENA Features							
cluster	**N**	**sleep**	**tCDS**	**ODS**	**AWC**	**CTC**	**CVC**	**noise**	**silence**	**distant**	**TV**	**meaningful**
4	2,041	**0.64**	0.22	0.14	3.0	0.1	0.5	0.01	*0.85*	0.08	0.02	0.03
5	142	**0.53**	0.30	0.17	3.4	0.1	0.9	*0.63*	0.12	0.16	0.04	0.04
6	1,256	0.00	**0.73**	0.27	54.6	*3.8*	*9.5*	0.02	0.27	0.25	0.01	*0.45*
1	3,450	0.01	**0.60**	0.39	22.0	1.1	4.8	0.03	0.37	0.33	0.03	0.25
7	1,485	0.01	0.33	**0.66**	*76.1*	1.4	2.6	0.01	0.21	0.33	0.03	*0.42*
2	3,475	0.04	0.45	**0.51**	13.6	0.4	1.8	0.03	0.17	*0.66*	0.02	0.12
3	1,087	0.27	0.28	**0.45**	7.3	0.2	0.7	0.02	0.15	0.07	*0.69*	0.06

Note: Bolded numbers correspond to clusters that included the highest proportion of segments classified most frequently as sleep, tCDS, and ODS, respectively. Italicized values indicate maximum cluster means of each LENA feature. AWC, CTC, CVC are automated counts per 5-minute segment, normalized to be rates (counts per minute). Values for noise, silence, distant, TV, and meaningful are proportions of each per 5-minute segment.

Classifying Sleep Segments

We attempted to build a classifier to automatically distinguish sleep from awake segments using only automatically-generated LENA features. All counts and durations of time were normalized to per-minute values (i.e., divided by segment duration). Although we experimented with simpler classification algorithms (e.g., decision trees and random forests; Bang et al., 2022), ultimately the best performance was achieved with XGBoost (eXtreme Gradient-Boosted trees; (Chen & Guestrin, 2016), a state-of-the-art algorithm that trains a cascade of decision trees successively on subsets of the data, upweighting the segments that were misclassified by earlier decision trees.[2] It

[2] XGBoost takes an MxN matrix of M training samples (5-minute segments, in our case) of N numeric features (scaled LENA metrics, here), and iteratively constructs a set of decision trees that aim to predict the given binary classes (e.g., sleep / not-sleep; or CDS / non-CDS), where each new tree focuses more on the data points that were misclassified by prior trees.

should be noted that XGBoost does not work well in some domains (e.g., it does not appear to work well for object recognition in images, Ohn-Bar & Trivedi, 2016), and tree-based methods in general do not extrapolate well beyond the range of feature values in the training set. Thus, it is important to thoroughly test via cross-validation, and to have a large and diverse training set to improve generalizability. An XGBoost classifier was trained using the xgboost R package (v1.7.5.1; Chen & He, 2023) to distinguish segments when the target child was asleep from those when they were awake, mirroring the first step that researchers could take when manually cleaning a LENA dataset. We trained the model using 5-fold cross-validation on 90% of the dataset (11,642 of 12,936 segments) and then tested the model on the remaining 10% held-out data (1,294 segments).

Results for the held-out data of the cross-validated classifier are shown in Figure 5. We illustrate the Receiver Operating Characteristic (ROC) curve, which depicts the performance of the classifier on sensitivity vs. specificity given all discrimination threshold values. On the left, the ROC curve reflects an overall ratio of sensitivity (y-axis) to specificity (x-axis) that was quite high, an Area Under the Curve (AUC) > .95, on the held-out test segments, with an accuracy of 0.945.[3] One limitation of XGBoost is that it does not enable simple visualizations, e.g., a decision tree, of how classifications are made. However, the feature importance measure can be used to assess which features were most informative in the ensemble of boosted trees. Shown in the right-hand panel of Figure 5, the amount of meaningful speech was the most important feature for classifying sleep segments, followed by the amount of silence, the number of child vocalizations, and distant speech.

A final sleep classifier was trained using all of the data (12,936 segments; 1,879 sleep segments, 11,057 awake), resulting in a classifier with superior performance to the cross-validated classifier (AUC = 0.985; see Appendix A for additional details). This sleep classifier has been made accessible for other researchers in a web app.[4]

Classifying tCDS vs. ODS Segments

We turn now to the more challenging task of building a classifier to automatically distinguish tCDS from ODS segments. We trained an XGBoost classifier on LENA features to distinguish tCDS segments from all other segments (ODS and split segments). First, we removed the 1,879 human-coded segments during which children were asleep (assuming they would be removed manually or by the sleep classifier). We then reclassified the 1,012 "split" segments which human coders judged to be 50% ODS and 50% tCDS as ODS, resulting in a total of 5,239 ODS segments and 5,818 tCDS segments

[3] To test whether the classifier was overfitting to characteristics of particular segments, we trained a 5-fold cross-validated version on 80% of the children, leaving out data from 20% of the children (n=30) in each fold. This classifier achieved very similar performance (AUC = 0.95; average test accuracy = 0.95), suggesting that the classifier will generalize to new children from similar samples. ROC curves for this analysis are shown in Appendix F.

[4] https://kachergis.shinyapps.io/classify_cds_ods/

(58% tCDS) when children were awake. The purpose of the classifier is thus to distinguish periods with >50% tCDS from segments that were at least 50% ODS, after removing periods of sleep. A random 90% of the awake data (9,951 out of 11,057 segments) was used to train the classifier, and the remaining 10% served as the held-out test set (1,106 segments) for evaluation.

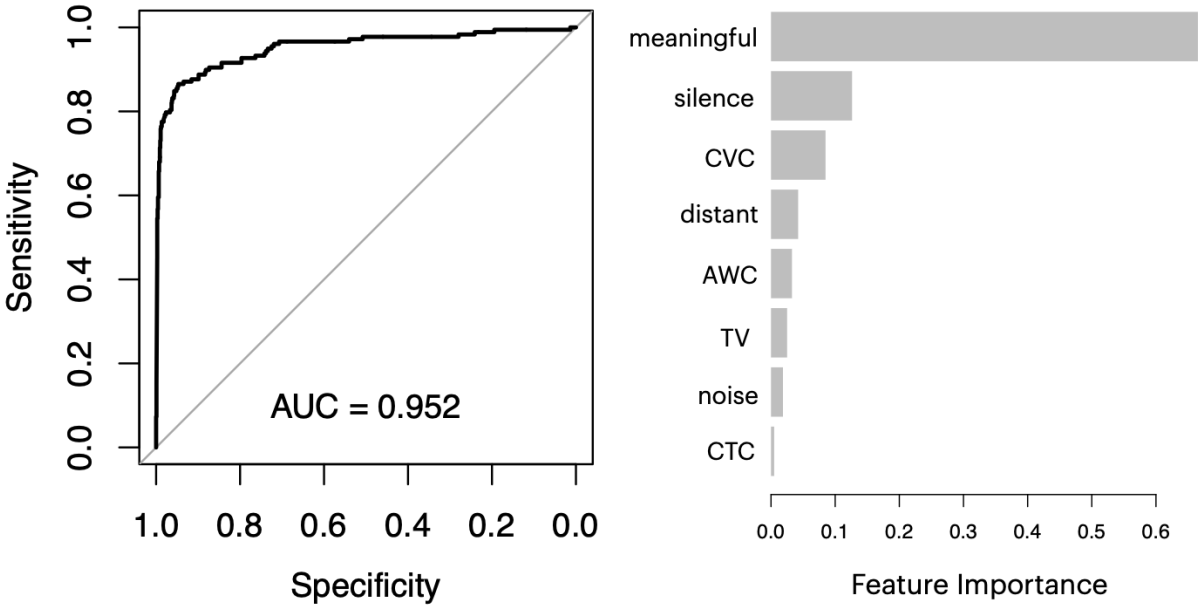


Figure 5. ***(left) The ROC curve for the sleep classifier for the 10% held-out test set. (right) Relative importance of the LENA features in the XGBoost sleep classifier trained on 90% of the data.***

The results are presented in Figure 6. As shown in the left-hand panel, when trained on 90% of the segments, the XGBoost classifier achieved moderate overall classification performance (AUC = 0.719), with an overall accuracy of 0.674 on the held-out data.[5] As shown in Figure 6 (right), the four most important features were the duration of silence, CTC, AWC, and meaningful speech.

A final XGboost classifier was trained with all 11,057 segments in order to offer the best chance for generalization to new data with similar samples, though there is no guarantee of similar performance for families and settings dissimilar to the present dataset. This final classifier's performance is shown in the Appendix Figure B1 and in

[5] To ensure that the classifier was not overfitting to these segments, we also trained a cross-validated version on 80% of the children, leaving out data from 20% of the children (n = 30) in each fold. This classifier achieved approximately the same performance (AUC = 0.73; average test accuracy = 0.66), suggesting that the classifier will generalize similarly well to data from additional children (see Appendix F for ROC curves). We also investigated including demographic and time of day features in the classifier, but found that inclusion of these features resulted in overfitting (i.e., poorer performance on held-out data).

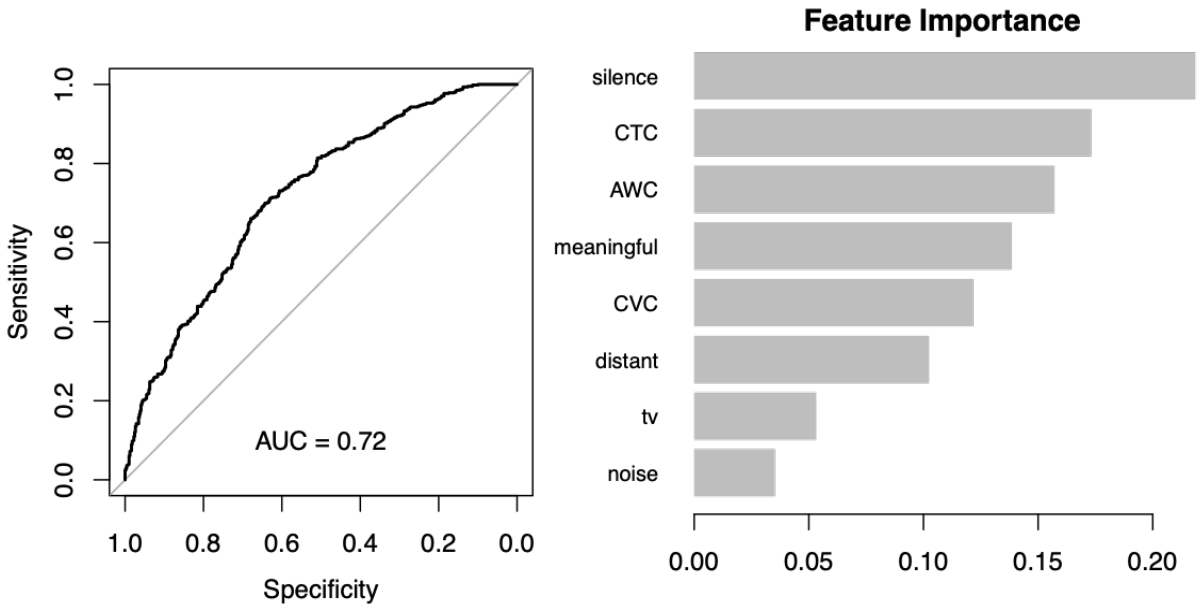


Figure 6. ***(left) ROC curve of the tCDS/ODS classifier for the 10% held-out test set and (right) relative importance of the LENA features in XGboost classifier trained on 90% of the data.***

the confusion matrices in Table 4. The AUC for this final classifier is much improved (AUC = 0.83), but performance on new data may be expected to be in-between the 90%-trained classifier and this higher value. This tCDS/ODS classifier is available for other researchers to use via a web app.[6]

Reliability Between the tCDS/ODS Classifier and a Human Rater

How does our classifier compare against human raters? Table 4a shows the confusion matrix for agreements (diagonal) and disagreements (off-diagonal) between the human raters (row) and the classifier's (columns) final binary predictions. The classifier correctly identified 80% of segments that humans rated as tCDS, as well as 70% of segments that humans rated as ODS. For comparison, Table 4b shows the confusion matrix for agreements (diagonal) and disagreements (off-diagonal) between two human raters. On average, human raters had 87% agreement for tCDS and 65% agreement for ODS. Thus, while tCDS agreements were slightly higher between two human raters and ODS agreements were slightly stronger between a classifier and a human rater, both confusion matrices indicate that ratings were similar whether comparing the classifier against a human rater or between two human raters. Sample-specific results can be seen in Appendix C.

[6] https://kachergis.shinyapps.io/classify_cds_ods/

Table 4. ***Confusion matrices.***

a) Human rater 1 (Gold Standard) vs. tCDS/ODS classifier

		Classifier		
		tCDS	ODS	Total
Human rater (Gold Standard)	tCDS	4641 (80% agreement)	1177	5818
	ODS	1554	3685 (70% agreement)	5239
	Total	6195	4862	11,057

Note: a) The diagonal (gray shading) indicates agreement between a human rater and the final XGboost classifier. b) The diagonal indicates agreement between two human raters. There were multiple individuals who served as first and second raters. For both tables, the percent agreement is calculated by dividing the number of agreements by the gold standard's total codes of the respective category.

b) Human rater 1 (Gold Standard) vs. Human rater 2

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS	190 (87% agreement)	28	218
	ODS	39	74 (65% agreement)	113
	Total	229	102	331

Links Between tCDS and Child Language Outcomes

One critical question is whether the tCDS/ODS classifier works sufficiently well to replicate results from studies with human-coded data. To test this, we used the Weisleder & Fernald (2013) dataset of 29 Spanish-speaking children whose caregivers completed the MacArthur-Bates Mexican Spanish CDI (Jackson-Maldonado et al., 2003) to assess vocabulary size when the children were 24 months. As illustrated in the left-hand panel of Figure 7, in this human-coded dataset, children who heard more tCDS at 19 months had significantly larger vocabularies at 24 months (r = .52, 95% CI = [.19, .75], p = .004). However, there was no significant association between the amount of ODS at 19 months and vocabulary size at 24 months (r = .25, p = .199).

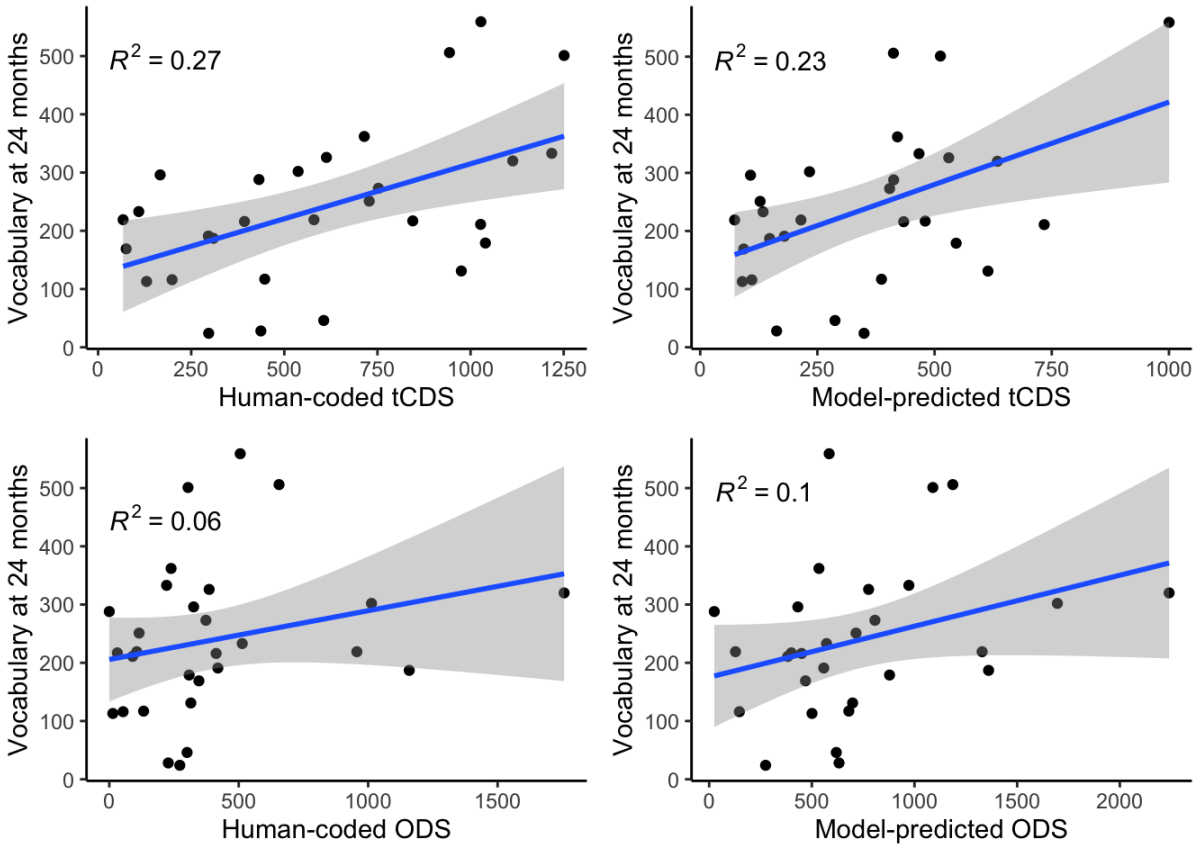


Figure 7. ***Scatterplots between human-coded or model-predicted tCDS or ODS at 19 months and children's later vocabulary sizes at 24 months. Note: Associations between vocabulary size and tCDS tokens are significantly positive, and of similar magnitude, whether human-coded (top, left) or model-predicted (top, right). Associations between vocabulary size and ODS tokens are not significant, but are of similar size, both for human-coded (bottom, left) and model-predicted (bottom, right) segments.***

We investigated these same correlations using the classifier's predictions of which segments were classified as tCDS vs. ODS. As shown in the right-hand panel of Figure

7, as with the original manual annotations, children who heard more tCDS at 19 months had significantly larger vocabularies at 24 months (r = .48, 95% CI = [.14, .72], p = .008), while the relation between the amount of ODS and vocabulary size was smaller, and did not achieve statistical significance by standard conventions (r = .32, 95% CI = [-.06, .61], p = .094). Notably, the pattern of the strength of the correlations are similar between the human-coded and model-predicted classifications, suggesting that the classifier is an effective tool for this purpose. To test whether this result was due to the inclusion of the Weisleder & Fernald dataset in the classifier's training set, we trained a classifier excluding this dataset, and found similar a pattern of results (tCDS vs. vocabulary r = .44, 95% CI = [.08,.69], p = .018; ODS vs. vocabulary r = .33, 95% CI = [−.04,.62], p = .082).

Leveraging Classifier Confidence

Although the classifier's binary performance is significantly above chance, there is substantial room for improvement, and thus we explored a more fine-grained measure of performance to determine whether some segments should be further examined by human coders. The source of XGboost's binary distinction between tCDS and ODS is actually a probability of tCDS in the range of [0,1], thresholded at 0.5 (i.e., if Pr(tCDS) > 0.5, a segment is classified as tCDS; otherwise it is classified as ODS). Figure 8 shows a histogram of classifier ratings (Pr(tCDS)) for all 11,057 awake segments in our full sample, color-coded by the classifications given by human listeners, with a dashed line indicating the threshold used for binary classification. Notably, there is significant overlap of the two distributions: there are many tCDS segments that (to the classifier) resemble and are thus confusable with ODS segments, and vice-versa. Of the 3,136 segments that were classified as tCDS with what could be considered to be low-confidence (0.4 < Pr(tCDS) < 0.6), 49% of them were judged by human coders to be tCDS. In contrast, a larger fraction of the segments classified with high confidence by the classifier agree with the human coder classification: for example, 89% of the 2,884 segments rated as Pr(tCDS) > 0.7 were judged to be tCDS by human coders, and 88% of the 2,196 segments rated as Pr(tCDS) < 0.3 were judged to be ODS by human coders. Thus, the probability of a segment being classified as tCDS could be used by researchers to make decisions about future coding or analysis, a point we return to in the discussion.

General Discussion

Our study suggests that a combination of automatically-generated measures of children's speech environments from LENA can be used to identify periods of sleep, tCDS, and ODS in daylong audio recordings, thus facilitating investigation of potentially meaningful sources of variation in young children's speech environments. We discuss our five main insights in turn.

First, we found differences in how the commonly-used, core frequency measures from LENA (AWC, CTC, and CVC) predicted the probability of a 5-minute segment being classified as containing primarily target-child-directed versus other-directed

speech. Our preliminary analyses indicated that segments with higher AWC relative to a family's mean were more likely to be judged by humans as having primarily other-directed speech. Frequency measures of CTC and CVC resulted in the opposite prediction, where segments with higher values relative to a family's mean were more likely to be judged as having predominantly target-child-directed speech. These findings suggest that periods of speech directed to a target child are defined by relatively lower rates of adult words and relatively higher rates of conversational turns and child vocalizations. This is consistent with the finding that adults often use a slower speech-rate when talking with children and that target-child-directed speech is more likely to elicit vocalizations from the child than other-directed speech. This finding also suggests that one reason some studies have found LENA's CTC measure to be a better predictor of child language outcomes than AWC (Gilkerson et al., 2018; Romeo et al., 2018) may be that high CTC is a better indicator of periods with target-child-directed speech than is AWC.

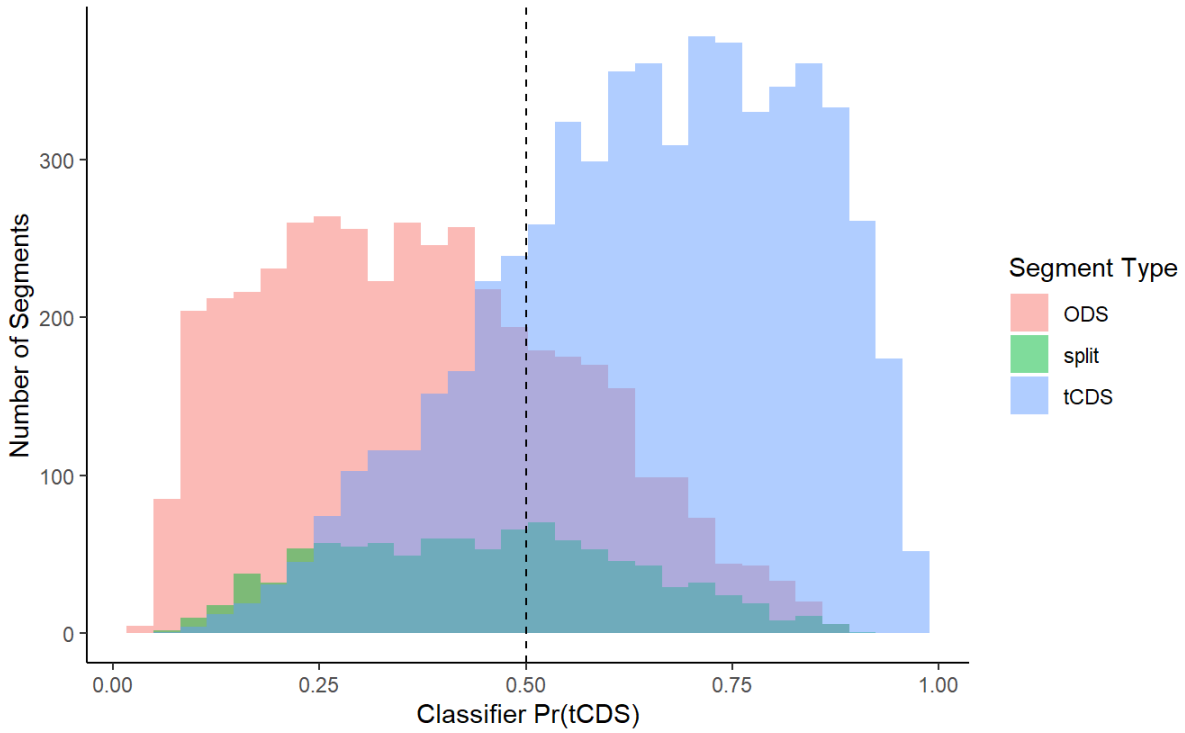


Figure 8. ***Histogram of classifier Pr(tCDS) for each segment, colored by human-coded segment type. Note: Dashed line indicates the threshold for binary classification: segments to the right were human-coded as tCDS (blue), while those to the left were human-coded as ODS (pink). Note that 'split' segments (green), which human coders found to be a mixture of both tCDS and ODS, were also given less decisive ratings of Pr(tCDS) by the classifier. The purple area indicates the overlap between tCDS and ODS regions.***

Second, a much more complex picture arose when including both LENA frequency

and duration measures in cluster analyses. While some distinct features characterized different audio environments, there was also a high degree of overlap across clusters. For example, as expected, clusters with more sleep segments were characterized by the lowest rates of AWC, CTC, and CVC. However, one sleep cluster was characterized by more silence, while the other was characterized by more noise. This aligns with anecdotal reports by human coders that periods of sleep sometimes involved what appeared to be fans or sound machines, sounds which were likely categorized as "noise" by LENA. Baby snores, which also sometimes occurred during periods of sleep, could also have been categorized as "noise" by LENA. In contrast, those clusters that were likely to be tCDS were characterized by the highest averages of CTC and CVC, but were more mixed with regards to AWC. Of clusters likely to be ODS, one cluster consisted of the highest average AWC, while the others had lower CTC and CVC rates, or longer durations of distant speech and TV. Thus, we observed multiple ways in which features were combined in clusters of predominantly sleep, tCDS, and ODS. Moreover, in no cases were sleep, tCDS, or ODS associated with only one cluster or configuration of features. Future work might fruitfully examine in more detail the potential differences between segments in different cluster types. For example, are segments in some clusters associated with different types of language interactions and/or activities than other segments?

Third, we found a high degree of success in training a classifier to identify periods of sleep in our dataset. Consistent with the multifaceted nature of clusters defined by more sleep, the classification was not simply due to periods of silence. The classifier mostly relied on the duration of 'meaningful' speech, followed by the duration of silence, and the number of vocalizations by the target child. This suggests that, at least among English- and Spanish-speaking families in the U.S., periods in which the target child is asleep vs. awake could be reliably identified from characteristics of the audio environment and shows advantages of considering multiple features of those environments.

Fourth, we found moderate success in training an XGBoost classifier to distinguish periods of tCDS versus ODS in our dataset. We found moderate sensitivity and specificity on the full dataset and a slightly smaller AUC on the held-out test segments. The feature importance list illustrated the average gain in our prediction of tCDS versus ODS, highlighting many features (meaningful speech, AWC, CTC, and silence) that also emerged in our cluster analysis. Reliability between two trained human raters suggests that even when individuals undergo training and interpret all available information in the auditory environment, there is variability across samples and there may be a ceiling of 'good enough' reliability. The moderate success of the classifier in terms of sensitivity and specificity, as well as performance seen in the confusion matrices, were similar to that of two human raters. This suggests that the level of accuracy achieved by the classifier may be a reasonable goal given the complexities of the speech environment. The superior performance of the classifier relative to analyses that were limited to individual predictors (i.e., the logistic regressions presented in our first analysis) suggests that human classifications of target-child-directed and

other-directed speech rely on nuanced distinctions that take into account combinations of features in the audio environment (e.g., low silence with high CTC and moderate AWC), as well as features of the environment not captured by these measures (e.g., semantic content).

Finally, we demonstrated that we could use model-derived predictions of tCDS and ODS to replicate associations between caregiver speech at 19 months and children's vocabularies at 24 months that were observed in previously published work in Spanish-speaking families in the U.S. (Weisleder & Fernald, 2013). We examined these correlations to test the performance of the classifier and not as an extension of the original study. The model-predicted classifications revealed, as observed with human-coding, that variability in speech to target children was positively and significantly correlated with children's later vocabularies, whereas this link was not statistically significant when using model-derived predictions of adult speech was directed to others.

Suggested Uses of the Classifier

We constructed a web app (https://kachergis.shinyapps.io/classify_cds_ods/) deploying the final XGboost classifiers for both sleep and tCDS/ODS, so that other researchers with daylong LENA recordings can easily use it on their datasets. However, it is important to note that this app has only been trained with data from U.S. families; thus, for researchers with populations dissimilar to those studied here, we recommend checks for the reliability of the classifier against human listeners (see Limitations below). Additionally, research on the generalizability of the classifier to new samples deserves separate attention, especially when considering which variables are theoretically motivated and logistically possible under different circumstances.

For those with LENA data, use of this web app may facilitate specifying the amount of speech directed to target children and speech directed to others. First, the sleep classifier can automate one laborious step of 'cleaning' daylong LENA recordings with a reasonably high degree of reliability. Second, the tCDS/ODS classifier could also be used to reduce the significant hours of manual labor required for coding periods of target-child-directed or other-directed speech. We have found that the classifier's per-segment probability of tCDS matches well with the uncertainty of human coders (e.g., the 50/50 "split" segments were classified as ~50% probability of being tCDS).

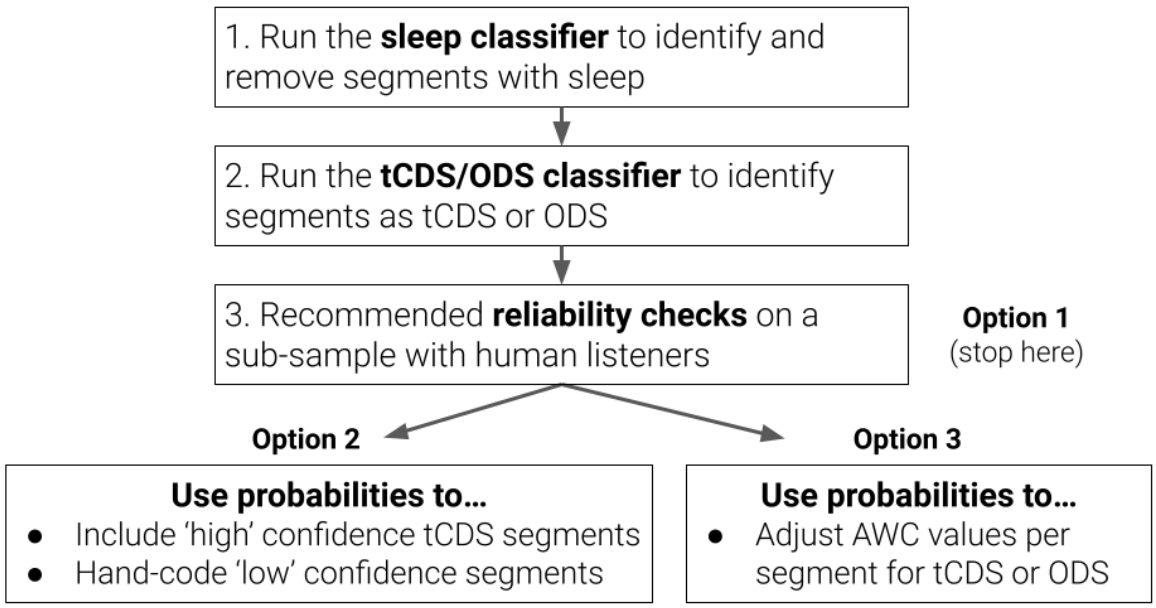
We suggest three potential workflows for using the classifier (Figure 9). Option 1 is to first run the sleep classifier to exclude periods when the child is sleeping and thus less likely to learn from the available speech, then running the tCDS/ODS classifier to identify binary judgements of segments considered as tCDS or ODS. Given that the classifier has been tested with a limited number of samples, we recommend reliability checks on a sub-sample of data with human listeners (if approved by ethics committees). To facilitate this, we provide our interrater reliability protocol for training human listeners, as well as the coding protocols from the original studies

(https://osf.io/qcj6r/). These protocols could also help guide reliability checks to compare human vs. classifier judgements.

Option 2 could be to follow the same steps, but rather than use a binary tCDS or ODS judgment, use the probabilities of tCDS or ODS to identify 'high confidence' tCDS or ODS segments versus 'low confidence' tCDS or ODS segments; 'low confidence' segments could then be listened to and judged by human coders. Whichever values are chosen, it is recommended to choose values that are symmetric (e.g., Pr(tCDS) < 0.3 (i.e., ODS) and Pr(tCDS) > 0.7 (i.e., tCDS)), to limit the introduction of bias.

Option 3 is to use the classifier probabilities to estimate the number of AWC tokens of tCDS and ODS in each segment by computing expected values (see Appendix D for more explanation). For example, a segment with an AWC of 200 and a .7 probability of being tCDS would result in 140 adult words counted as tCDS and 60 words counted as ODS for that segment. Rather than binning segments based on a binary probability of the entire segment falling into the tCDS versus ODS category, each 5-minute segment would contribute some of its counts to both. See Appendix D for an application of this method to the Weisleder & Fernald (2013) data, which yielded similar associations with outcomes. It is important to note that higher tCDS probabilities may reflect more of a certain type of verbal interaction (e.g., one-on-one interactions in a quiet indoor setting) than other types of caregiver-child interactions (e.g., playing outside where speakers may be further away from each other). Therefore, how probabilities are used should be considered with caution and transparently documented to better understand their utility and significance.

Figure 9. ***Potential workflows with the classifier.***



Limitations

While we included over 1,000 hours of data from 153 English- and Spanish-speaking families from varied socioeconomic backgrounds, our sample nevertheless represents a small subset of the variability that exists within English- and Spanish-speaking families in the U.S. and a tiny subset of the linguistic (e.g., different languages, multilingualism, signed vs. spoken language), cultural, and ecological variability in child-rearing environments around the world. For example, given the wide variability in infant sleep routines seen across families and countries throughout the world (Mindell et al., 2010), most of which are not represented in our training data, it is possible that the LENA features that characterize periods of sleep in our recordings will not generalize to recordings collected in very different contexts. In particular, all of the families in our studies lived in urban settings, and it is likely that the LENA features that characterize periods of sleep would differ for families in different settings (e.g., subsistence farming communities; Casillas et al., 2019, 2021). Similarly, given the wide variability in ways of interacting with children observed across sociocultural settings, it is possible that the features that differentiated tCDS from ODS in our sample of English- and Spanish-speaking families in California will not generalize to other contexts. Further validation studies are critical to understand whether our classifiers can generalize to new languages and communities (Cristia et al., 2021). Other studies that have coded tCDS vs. ODS in various other languages and contexts (Tseltal in a Mayan village: Casillas et al., 2019; Yélî Dnye in a Papuan community: Casillas et al., 2021; Spanish in Argentina: Rosemberg et al., 2020; Sesotho in South Africa and French in France: Loukatou et al., 2022) have done this in different ways (e.g., utterance-level coding vs. global binary judgements of tCDS or ODS). Thus, at the moment, our classifier cannot be applied to these data. Additionally, while our classifier is open-source, LENA software is not; thus, the ability to use this classifier requires a substantial cost to purchase the LENA recorders and software. Future work should compare whether our classifiers can be used with open-source speech algorithms (e.g., ALICE; Räsänen et al., 2021) to achieve similar performance. Finally, while the classifier can facilitate identification of periods of sleep, tCDS, and ODS in daylong audio recordings, this automated method does not reveal the specific acoustic, linguistic, or interactional features that differentiate between these speech contexts. Thus, it is far from replacing the need for human annotation and transcription and more research is needed to better explain how children learn from the language(s) to which they are exposed.

Conclusion

These findings suggest exciting opportunities for advancing our understanding of how children learn from the available speech in their environment. We were able to train and validate two automated classifiers using LENA-based measures to identify periods of sleep and to distinguish between periods of tCDS versus ODS. This work has the potential to significantly reduce the time-consuming process of identifying periods of directed speech to target children from the rich and naturalistic information collected with daylong recordings. In this way, the progress that we have

made here can facilitate future research seeking to illuminate questions about the relations between target-child-directed and other-directed speech on child outcomes and/or about the features of child-directed speech across linguistically- and culturally-diverse communities. We hope this adds to existing methods to explore shared and different features of target-child- and other-directed speech so we can better understand how different children across diverse communities acquire and develop their language skills.

References

Akhtar, N., Jipson, J., & Callanan, M. (2001). Learning words through overhearing. *Child Development, 72*(2), 416–430. https://doi.org/10.1016/S0163-6383(98)91471-0

Bakeman, R. (2023). KappaAcc: A program for assessing the adequacy of kappa. *Behavior Research Methods, 55*(2), 633-638. https://doi.org/10.3758/s13428-022-01836-1

Bang, J. Y., Mora, A., Munévar, M., Fernald, A., & Marchman, V. A. (2022). *Time to talk: Multiple sources of variability in caregiver verbal engagement during everyday activities in English- and Spanish-speaking families in the U.S*. PsyArXiv. https://doi.org/10.31234/osf.io/6jzwg

Busch, T., Sangen, A., Vanpoucke, F., & Van Wieringen, A. (2018). Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods, 50*(5), 1921–1932. https://doi.org/10.3758/s13428-017-0960-0

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., Durrant, S., Fennell, C. T., Fiévet, A.-C., Frank, M. C., Gampe, A., Gervain, J., Gonzalez-Gomez, N., Hamlin, J. K., Havron, N., Hernik, M., Kerr, S., Killam, H., Klassen, K., … Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920974622. https://doi.org/10.1177/2515245920974622

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods, 48*(3), 1109–1124. https://doi.org/10.3758/s13428-015-0634-8

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language, 48*(4), 792–814. https://doi.org/10.1017/S0305000920000549

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2023). *xgboost: EXtreme Gradient Boosting* (1.7.5.1). https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

Crago, M. B., Allen, S. E. M., & Hough-Eyamie, W.P. (1997). Exploring innateness through cultural and linguistic variation. In M. Gopnik (Ed.), *The inheritance and innateness of grammars* (pp. 70–90). New York City, NY, USA: Oxford University Press.

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

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & 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

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., de Barbaro, K., Bang, J. Y., & Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods, 52,* 1951-1969. https://doi.org/10.3758/s13428-020-01365-9

Cychosz, M., Villanueva, A., & Weisleder, A. (2021). Efficient estimation of children's language exposure in two bilingual communities. *Journal of Speech, Language, and Hearing Research, 64*(10), 3843–3866. https://doi.org/10.31234/osf.io/dy6v2

Dailey, S., & Bergelson, E. (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science, 25*(3), e13192. https://doi.org/10.1111/desc.13192

De Palma, P., & VanDam, M. (2017). Using automatic speech processing to analyze fundamental frequency of child-directed speech stored in a very large audio corpus. *IEEE Proceedings of IFSA-SCIS,* 1–6. https://doi.org/10.1109/IFSA-SCIS.2017.8023224

Ferjan Ramírez, N., Hippe, D. S., Braverman, A., Weiss, Y., & Kuhl, P. K. (2023). A comparison of automatic and manual measures of turn-taking in monolingual and bilingual contexts. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02127-z

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language, 32*(3), 385–397. https://doi.org/10.1177/0142723711433584

Gilkerson, J., & Richards, J. A. (2020). *A guide to understanding the design and purpose of the LENA system*. LENA Foundation. https://www.lena.org/wp-content/uploads/2020/07/LTR-12_How_LENA_Works.pdf

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4), e20174276. https://doi.org/10.1542/peds.2017-4276

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., Harnsberger, J., & Topping, K. (2015). Evaluating Language Environment Analysis System Performance for Chinese: A Pilot Study in Shanghai. *Journal of Speech, Language, and Hearing Research, 58*(2), 445–452. https://doi.org/10.1044/2015_JSLHR-L-14-0014

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2020). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, 1–12. https://doi.org/10.1101/2020.04.09.032995

Hoff, E., Burridge, A., Ribot, K. M., & Giguere, D. (2018). Language specificity in the relation of maternal education to Bilingual Children's vocabulary growth. *Developmental Psychology, 54*(6), 1011–1019. https://doi.org/10.1037/dev0000492

Jackson-Maldonado, D., Thal, D., J., & Fenson, L. (2003). *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual*. Brookes Publishing.

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2021). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior Research Methods, 53*(1), 113–138. https://doi.org/10.3758/s13428-020-01419-y

Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language, 49*(6), 1173–1192. https://doi.org/10.1017/S0305000921000623

Marchman, V. A., Bermúdez, V. N., Bang, J. Y., & Fernald, A. (2020). Off to a good start: Early Spanish-language processing efficiency supports Spanish- and English-language outcomes at 4½ years in sequential bilinguals. *Developmental Science, 23*(6), e12973. https://doi.org/10.1111/desc.12973

McGillion, M. L., Herbert, J. S., Pine, J. M., Keren-Portnoy, T., Vihman, M. M., & Matthews, D. E. (2013). Supporting early vocabulary development: What sort of responsiveness matters? *IEEE Transactions on Autonomous Mental Development, 5*(3), 240–248. https://doi.org/10.1109/TAMD.2013.2275949

Mendoza, J. K., & Fausey, C. M. (2021). Quantifying everyday ecologies: Principles for manual annotation of many hours of infants' lives. *Frontiers in Psychology, 12*, 710636. https://doi.org/10.3389/fpsyg.2021.710636

Mindell, J. A., Sadeh, A., Wiegand, B., How, T. H., & Goh, D. Y. T. (2010). Cross-cultural differences in infant and toddler sleep. *Sleep Medicine, 11*(3), 274–280. https://doi.org/10.1016/j.sleep.2009.04.012

Ochs, E., & Schieffelin, B. (1984). Language acquisition and socialization: Three developmental stories and their implications. In R. Schweder, & R. Levine (Eds.) *Culture theory: Essays on mind, self, and emotion* (pp. 276–322). Cambridge University Press.

Ohn-Bar, E., & Trivedi, M. M. (2016). To boost or not to boost? On the limits of boosted trees for object detection. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3350–3355. https://doi.org/10.1109/ICPR.2016.7900151

Quigley, J., Nixon, E., & Lawson, S. (2019). Exploring the association of infant receptive language and pitch variability in fathers' infant-directed speech. *Journal of Child Language, 46*(04), 800–811. https://doi.org/10.1017/S0305000919000175

Räsänen, O., Kakouros, S., & Soderstrom, M. (2018). Is infant-directed speech interesting because it is surprising? – Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition, 178,* 193–206. https://doi.org/10.1016/j.cognition.2018.05.015

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods, 53*(2), 818–835. https://doi.org/10.3758/s13428-020-01460-x

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Rosemberg, C. R., Alam, F., Audisio, C. P., Ramirez, M. L., Garber, L., & Migdalek, M. J. (2020). Nouns and verbs in the linguistic environment of Argentinian toddlers: Socioeconomic and context-related differences. *First Language, 40*(2), 192–217. https://doi.org/10.1177/0142723719901226

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A. S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., … Zafeiriou, S. (2017). The INTERSPEECH 2017 Computational paralinguistics challenge: Addressee, cold & snoring. *Interspeech 2017,* 3442–3446. https://doi.org/10.21437/Interspeech.2017-43

Schuster, S., Pancoast, S., Ganjoo, M., Frank, M. C., & Jurafsky, D. (2014). Speaker-independent detection of child-directed speech. *2014 IEEE Spoken Language Technology Workshop (SLT),* 366–371. https://doi.org/10.1109/SLT.2014.7078602

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science, 15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy, 14*(6), 654–666. https://doi.org/10.1080/15250000903263973

Snow, C. E. (1977). Mothers' speech research: From input to interaction. In C. Snow, & C. A. Ferguson (Eds.), *Talking to children* (pp. 31–49). Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE, 8*(11). https://doi.org/10.1371/journal.pone.0080646

Solomon, O. (2011). Rethinking baby talk. In A. Duranti, E. Ochs, & B. B. Schieffelin (Eds.), *The Handbook of Language Socialization* (1st ed., pp. 121–149). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444342901.ch5

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech, 65*(1), 3–27. https://doi.org/10.1177/0023830920979016

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science, 23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Tamis-LeMonda, C. S., Song, L., Leavell, A. S., Kahana-Kalman, R., & Yoshikawa, H. (2012). Ethnic differences in mother-infant language and gestural communications are associated with specific skills in infants: Mother-infant communications. *Developmental Science, 15*(3), 384–397. https://doi.org/10.1111/j.1467-7687.2012.01136.x

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore, P. Dunham, J. Philip (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE, 11*(8), e0160588. https://doi.org/10.1371/journal.pone.0160588

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/10.3399/096016407782317928

Xu, D., Yapanel, U., Gray, S., & Baer, C., T. (2009). *The LENA Language Environment Analysis System: The Interpreted Time Segments (ITS) File*. LENA Foundation. https://www.lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf

Yurovsky, D. (2018). A communicative approach to early word learning. *New Ideas in Psychology, 50*, 73–79. https://doi.org/10.1016/j.newideapsych.2017.09.001

Data, code and materials availability statement

All data and analysis code for the app are available here: https://github.com/kachergis/classify_cds_ods. All analysis code for the current manuscript are available here: https://github.com/kachergis/tCDS_nap_classifier_paper. Our protocol to determine human to human interrater reliability are available here: https://osf.io/qcj6r/

Ethics statement

Ethics approval was obtained from the Stanford University Institutional Review Board. All families provided informed consent prior to their participation in the study.

Authorship and Contributorship Statement

All authors contributed to conceptualization of the present study. J.B. (logistic regressions and reliability) and G.K. (classifiers) contributed to analyses. J.B., A.W., and V.M. supervised and analyzed original coding of LENA data across the five samples. J.B. and G.K. contributed to the original draft preparation. All authors contributed to writing, reviewing, and editing of the final manuscript.

Acknowledgements

We are especially grateful to the families for their contribution to this research. We would also like to thank Anne Fernald for supporting the studies that enabled this work, the Language Learning Lab staff for their tireless work in hand-coding these data, and the members of the Language and Cognition lab at Stanford and the LangVIEW consortium for their thoughtful comments and suggestions. This work was supported by grants from the National Institutes of Health (R01 HD42235, HD092343, HD069150), the Schusterman Foundation, the W.K. Kellogg Foundation, the David and Lucile Packard Foundation, and the Bezos Family Foundation to Anne Fernald, the National Institutes of Health (2R01 HD069150) to Heidi Feldman, the National Institutes of Health (R21 DC018357) and a Elizabeth Munsterberg Koppitz Child Psychology Graduate Student Fellowship from the American Psychological Foundation to Adriana Weisleder, and a Postdoctoral Support Award from the Stanford Maternal and Child Health Research Institute to Janet Bang. This work is an extension of work published in the Proceedings of the 46th annual Boston University Conference on Language Development (Bang, Kachergis, Weisleder, & Marchman, 2022).

Appendices

Appendix A. Final Sleep Classifier

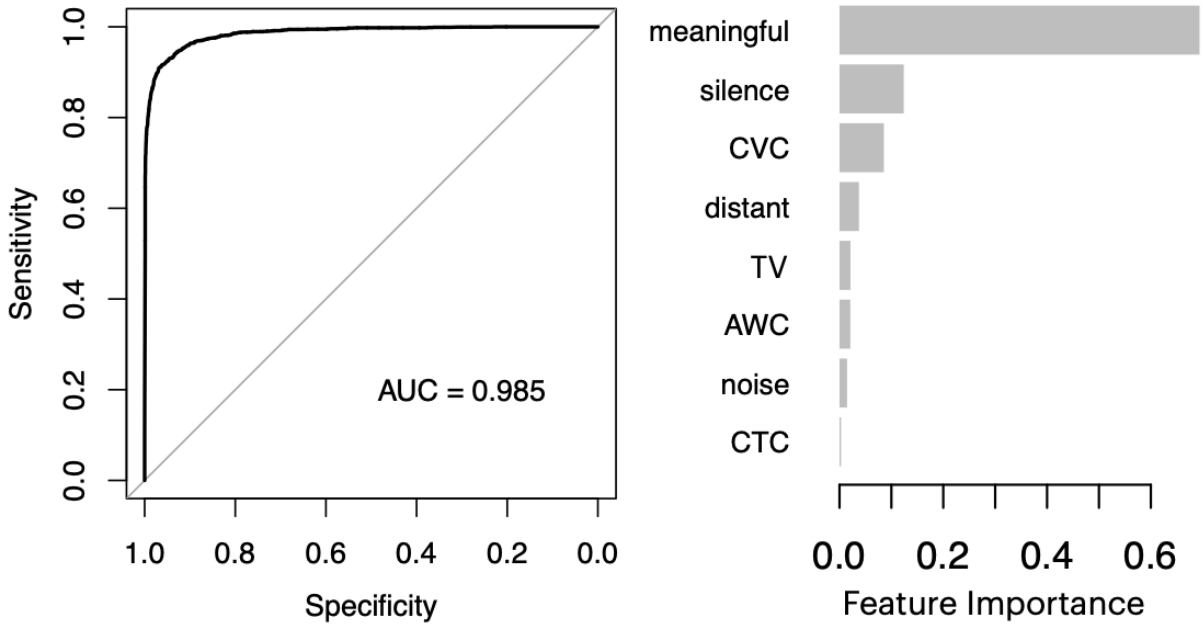


Figure A1. ***The final XGBoost sleep classifier, trained on the entire dataset, has a slightly higher AUC than the cross-validated classifier had on held-out data (Figure 5). The relative feature importances are quite similar to the held-out data classifier, although TV became slightly more important than AWC in the final classifier.***

Appendix B. Final tCDS/ODS Classifier

A final XGboost classifier was trained on the entire set of tCDS and ODS segments, and the results of this classifier are shown in Figure B1. The feature importances are similar to those in the classifier trained on 90% of the data, except that there is more reliance on AWC and slightly less on CTC in the final classifier. The AUC is also much improved.

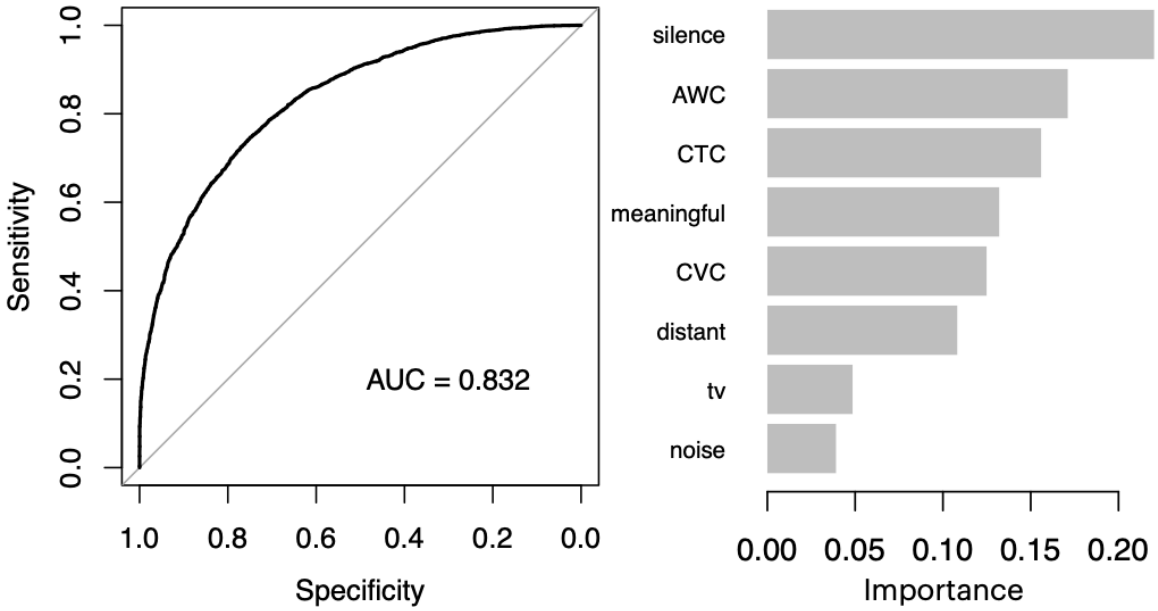


Figure B1. ***(left) ROC curve of the tCDS/ODS classifier and (right) relative importance of the LENA features in the final XGboost classifier trained on all 11,057 segments.***

Appendix C. Confusion Matrices Between Two Human Raters When Examining Interrater Reliability per Sample

Note that the diagonal (gray shading) indicates agreement between two human raters. For all tables, the percent agreement is calculated by dividing the number of agreements by the gold standard's total codes of the respective category.

Table C1. ***Sample 1***

		Human rater 2		
		tCDS	ODS	Total
Human Rater 1 (Gold Standard)	tCDS (rater 1)	46 (92% agreements)	4	50
	ODS (rater 1)	5	4	9

		(44% agreements)	
Total	51	8	59

Table C2. ***Sample 2***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	28 (100% agreements)	0	28
	ODS (rater 1)	11	17 (61% agreements)	28
	Total	39	17	56

Table C3. ***Sample 3***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	37 (84% agreements)	7	44
	ODS (rater 1)	9	25 (74% agreements)	34
	Total	46	32	78

Table C4. ***Sample 4***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	32 (76% agreements)	10	42
	ODS (rater 1)	1	21 (95% agreements)	22
	Total	33	31	64

Table C5. ***Sample 5***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	47 (87% agreements)	7	54
	ODS (rater 1)	13	7 (35% agreements)	20
	Total	60	14	74

Appendix D. Using Classifier Probabilities to Estimate tCDS and ODS Tokens

Future research may benefit from using the classifier-estimated probability of each segment being tCDS in two ways: 1) to reduce the amount of human coding (e.g, by only listening to the low-confidence segments), and 2) to estimate the number of tCDS and ODS tokens in each segment. First, given the high accuracy of the classifier for high-confidence classifications (~92% for Pr(tCDS) > 0.7 (i.e., tCDS), and 77% Pr(tCDS) < 0.3 (i.e., ODS)), one could use the classifier predictions for these segments, while potentially choosing to code the remaining low-confidence segments by hand. For the present dataset, this would have reduced the time needed to code the segments by 46%. For researchers primarily interested in segments that are likely to be primarily tCDS, it may be justified to disregard the likely ODS segments (e.g., Pr(tCDS) < 0.3; ~20% of our dataset). Determining what criterion to use requires careful consideration of the goals of the research, but there may be additional utility in leveraging the classifier's immediate, fine-grained judgments to support human rating for more difficult segments.

Moreover, the classifier's probability rating for each segment could be interpreted as an estimated proportion of the segment's tCDS (vs. ODS) content, and researchers could use the estimated tokens of tCDS and ODS AWC to calculate an expected value of both tCDS and ODS tokens for each child. That is, if a given 5-minute segment with 100 adult words receives a rating of Pr(tCDS) = 0.75, then the expected number of tCDS tokens in that segment is Exp(tCDS) = 100*0.75 = 75, and Exp(ODS) = 100*(1-Pr(tCDS)) = 25 tokens. Using this more fine-grained measure of each segment's contents may provide a better signal, as compared to the binarized classification, which assigns each segment's AWC tokens to either tCDS or ODS. Whether a segment with a higher probability of tCDS actually contains more tCDS (and less ODS) is an empirical question, which we will indirectly address here by examining the relation between children's classifier-rated amount of experienced tCDS and ODS and their later vocabulary size using the data from Weisleder & Fernald (2013), as before. The correlation for Exp(tCDS) and vocabulary size at 24 months was r = .56 (t(27) = 3.53, 95% CI = [.25, .77], p = .001), which is somewhat higher than when using the binary tCDS/ODS

judgments, from either the classifier or the human raters. The correlation for Exp(ODS) and vocabulary size at 24 months was r = .35 (t(27 = 1.94, 95% CI = [-.02, .64], p = .06), roughly similar to what was found using the binary judgments. Another hint that the classifier's Pr(tCDS) rating may correspond to humans' confidence is that the majority (74%) of the 'split' segments identified by human raters had great uncertainty for the classifier: only 26% of these segments were given high-confidence ratings in the model (Pr(tCDS) < 0.3 or Pr(tCDS) > 0.7).

Appendix E. Testing Classifiers Using Sample-Level Cross-Validation

Given that these samples were collected over many years, with potential variation in populations and training of research assistants, we chose to test whether leaving contemporaneously collected samples out of the training set unduly influenced the performance of the sleep or tCDS/ODS classifiers. Table E1 shows the accuracy and AUC for sleep classifiers trained without each sample, showing that performance was fairly consistent (accuracy range: [0.945,0.970]; AUC range: [0.948,0.985]). Table E2 shows the results for tCDS/ODS classifiers trained without each sample, showing broadly similar performance (accuracy range: [0.628,0.708]; AUC range: [0.690,0.786]). It is worth noting that leaving out Sample 1 does somewhat decrease performance, and leaving out Sample 4 somewhat increases performance. Nonetheless, on balance we believe that including the full dataset gives the greatest chance of generalizing to new datasets.

Table E1. ***Sleep classifier results when respective samples are left out***

Sample-left-out	Accuracy of classifier without sample	Area Under Curve (AUC) without sample
Sample 1	0.970	0.985
Sample 2	0.945	0.955
Sample 4	0.948	0.966
Samples 3 and 5*	0.967	0.948

Table E2. ***tCDS/ODS classifier results when respective samples are left out***

Sample-left-out	Accuracy of classifier without sample	Area Under Curve (AUC) without sample
Sample 1	0.628	0.690
Sample 2	0.666	0.721
Sample 4	0.708	0.786
Samples 3 and 5*	0.683	0.723

Note: *Samples 3 and 5 were both coded at the 10-min level, and then split into 5-min segments to include in the classifier. We group them here to cross-validate the classifier.

Appendix F. Testing Classifiers Using Child-Level Cross-Validation

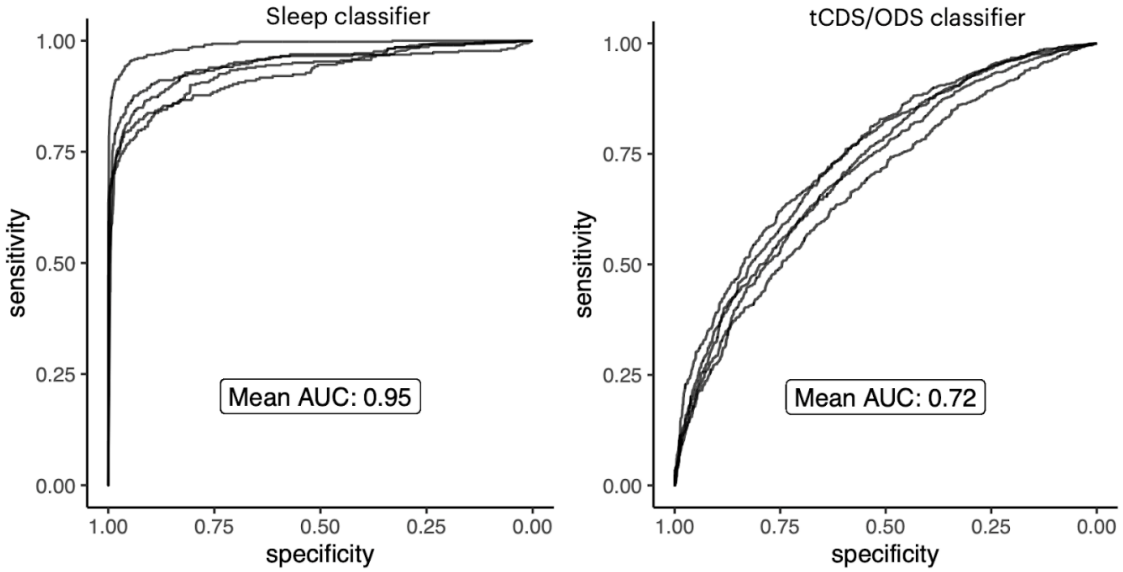


Figure F1. ***ROC curves for classifiers that exclude 20% of children (n = 30) in each training set, for sleep (left) and tCDS/ODS (right).***

License