Accuracy of the Language Environment Analysis (LENATM) System Segmentation and Metrics:
A Systematic Review

Alejandrina Cristia^a, Federica Bulgarelli^b, & Erika Bergelson^b

^a Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études

cognitives, ENS, EHESS, CNRS, PSL University

^b Psychology & Neuroscience, Duke University, Durham, North Carolina

Author note

This work was supported by a Trans-Atlantic Platform “Digging into Data” collaboration grant (ACLEW: Analyzing Child Language Experiences Around The World), with the support of Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW; in addition to ANR-17-CE28-0007, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and the National Endowment for the Humanities (HJ- 253479-17), as well as funding from the J. S. McDonnell Foundation (AC) and NIH Grant DP5-OD019812 (EB). The funders had no impact on this systematic review. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work and have no conflict of interests to disclose.

Correspondence concerning this article should be addressed to Alejandrina Cristia, 29, rue d'Ulm, 75005 Paris, France. E-mail: alecristia@gmail.com,

Abstract

Purpose: The Language Environment Analysis (LENATM) system provides automated measures facilitating clinical and non-clinical research and interventions on language development, but there are only a few, scattered independent reports of these measures' validity. The objectives of the current systematic review were to (1) Discover studies comparing LENATM output with manual annotation, namely accuracy of talker labels, as well as involving Adult Word Counts (AWC), Conversational Turn Counts (CTC), and Child Vocalization Counts (CVC); (2) Describe them qualitatively; (3) Quantitatively integrate them to assess central tendencies; and (4) Quantitatively integrate them to assess potential moderators.

Method: Searches on Google Scholar, PubMed, Scopus, and PsycInfo were combined with expert knowledge, and inter-article citations resulting in 238 records screened, and 73 records whose full-text was inspected. To be included, studies must target children under age 18 years and report on accuracy of LENATM labels (e.g., precision and/or recall), and/or AWC, CTC, or CVC (correlations and/or error metrics).

Results: A total of 33 studies, in 28 articles, were discovered. A qualitative review revealed most validation studies had not been peer-reviewed as such and failed to report key methodology and results. Quantitative integration of the results was possible for a broad definition of recall and precision (mean = 59% and 68% respectively, N = 12-13), for AWC (mean $r = .79$, N = 13), CVC (mean $r = .77$, N = 5), and CTC (mean $r = .36$, N = 6). Publication bias and moderators could not be assessed meta-analytically.

Conclusion: Further research and improved reporting are needed in studies evaluating LENA segmentation and quantification accuracy, with work investigating CTC being particularly urgent.

Accuracy of the Language Environment Analysis (LENATM) System Segmentation and Metrics: A Systematic Review

Over the past decade, there has been a sea change in how early childhood specialists think about and analyze the language environment that infants experience and contribute to. This has been driven in no small part by improving technology, which allows for longer, less obtrusive recordings, and automated analyses of recording contents. The Language Environment Analysis (LENATM, Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011) system is perhaps the most prominent ‘off the shelf’ system, providing researchers, clinicians, and early educators a snapshot of young children’s input and speech productions with virtually no effort. The system combines a small wearable device that records for up to 16h, with automated speech analyses of these long recordings, determining who is talking, how much, and when. These kinds of numbers, in turn, can be used to assess where a given child’s vocalizations fall relative to age-matched peers, and to guide interventions seeking to assess and increase children’s ‘language nutrition’ (Oller et al., 2010; e.g. D. L. Suskind et al., 2013). Two key advantages of the LENATM system are that it provides automated output without requiring any manual annotation, and that by measuring an entire day, it provides more ecologically valid data than can be gathered when infants come into a lab or clinic, or in brief home recording contexts (Bergelson et al., 2018; e.g., Tamis-LeMonda, Kuchirko, Luo, Escobar, & Bornstein, 2017).

Thus, the LENATM system has been an important addition to clinicians’, psychologists’, and linguists’ toolkits, as evidenced by a growing literature in both the basic science and intervention realms (Adams et al., 2018; Sosa, 2016; D. L. Suskind et al., 2016, 2013; Wood, Diehm, & Callender, 2016). A number of independent evaluations of the LENATM system have

also become public since validation of several LENATM metrics were published as part of LENATM Foundation reports (D. Xu, Yapanel, & Gray, 2009).

We aimed to summarize work comparing the LENATM automatic output against human annotation of the same data, for both basic and derived LENATM annotations. Regarding basic annotations, the system divides the audio into a few classes to decide who speaks when. The speech classes are: Female Adult, Male Adult, Key Child, Other Child. The non-speech classes are: Silence, TV and Electronic Noise, Undefined Noise, and Overlap (among any of the above). Additionally, LENATM can label a section as “far” versions of the other tags (e.g., Female Adult Far) when the model finds the section is equally likely to be Female Adult or Silence. Adopting the standard approach, we aimed to calculate classification accuracy as recall and precision. Details on how this was operationalized are below, but in short, recall indicates how accurate the system is in detecting an event (e.g., a given speech class) when the event occurs; and precision indicates how often the system is right when it says there was an event.

As for derived metrics, the LENATM system also provides quantitative information about how much talking is happening, i.e., Adult Word Count (AWC), Child Vocalization Count (CVC), and Conversational Turn Count (CTC). We sought to quantitatively integrate estimates of the accuracy of these metrics, which by virtue of being numeric (rather than categorical like speech class) can be evaluated using correlation coefficients (i.e., how much variance is shared across estimates provided by the LENATM system and by human coders) and relative error rates (i.e., how far off the LENATM system estimates and the human coders’ estimates are).

Additionally, we considered two sets of potential moderators. LENATM’s algorithms were originally trained on data from a sample of 1- to 42-month-old children learning North American

English (NAE), balanced in gender and stratified by maternal education. Specifically, the training used 135 hours of data from 309 separate recordings from as many infants, sampling a 30-minute section from each recording (Gilkerson, Coulter, & Richards, 2008).¹ We intended to look at accuracy as a function of several participant- and method-related moderators to assess the generalizability of LENATM output beyond the kind of data in the original training set. For *participant*-related moderators, we considered native language, matching status (i.e., whether the participants matched LENATM's training sample or not), and mean age and age range. We expected that studies that did not match LENATM's training sample (e.g., because it was a different language or an older mean age) would have *lower* reliability. Relatedly, wider age ranges may lead to *higher* reliability because the participants may be more diverse, and thus observations for each of them may be more different from each other.

For *methods*-related moderators, we considered how clips were selected and presented to human annotators, and the total duration of annotations cumulated across participants in each study (further details in the Methods section). These methodological factors were relevant because they affect the extent to which we can view validation results as biased by the same algorithm they intend to evaluate, and how those results may generalize to day-long data at large. For example, if human annotators know that the LENATM system classified a stretch of audio as

¹ Although other documents from the LENATM Foundation state this age range is 2 to 48 months (e.g., Richards, Gilkerson, Paul, & Xu, 2008), we use the range in the document that specifically reports on the transcriptions of the training and test sets used for training acoustic models for segmentation (Gilkerson et al., 2008).

“female adult”, it may take more effort for them to notice that this stretch of audio also contains some silence or noise or overlap with a different talker. In contrast, human annotators who did not have access to LENATM’s labels cannot be influenced by them. " As for total duration, we reasoned this may be a measure of data quantity, and could potentially point to evidence of biased reporting, if there were any. In many meta-analyses, an anti-correlation between data quantity and effect size (i.e., smaller studies have better results than bigger studies) is consistent with this kind of selective reporting. Following the same rationale, we reasoned that an anti-correlation of reliability and amount of data may be consistent with selective reporting.

Methods

All decision steps are documented in the supplementary materials (Cristia, Bulgarelli, & Bergelson, 2019a), which also provides access to code necessary to reproduce all analyses below.

The eligibility criteria for inclusion in the qualitative analyses of the systematic review were: (1) The LENATM device was worn by children aged 18 years or less, with no limits placed on native language, population sample, recording setting, or recording duration (which led to the inclusion of 2 studies at least partially based on very short recordings); (2) The LENATM audio was coded by a human annotator; (3) Accuracy was estimated via agreement for categorical decisions (e.g., speaker labels) and correlations and/or error rates for derived estimates (Adult Word Count, Child Vocalization Count, Conversational Turn Count). As a result of these deliberately widely inclusive criteria, studies were only excluded if there was no accuracy discussed.

To be included in the quantitative analyses, a paper needed to report quantitatively on one or more of the following metrics: precision and/or recall for categorical decisions; Pearson correlations on AWC, CVC, CTC; relative error rates for AWC, CVC, CTC; or means for LENATM and human AWC, CVC, CTC (so that we could derive relative error rates).

The information sources used to compose the initial list included suggestions by experts (authors of this work, as well as Melanie Soderstrom); three google scholar searches (example of a full keyword set: “LENATM, speech reliability, male-adult, female-adult, child”) and a google alert complemented with searches in Scopus, PubMed, and PsycInfo; as well as papers cited as containing an evaluation by any of the full papers inspected. We stopped including studies on 2019-09-10. The PRISMA flowchart is in Figure 1.

–INSERT FIGURE 1 ABOUT HERE –

Data were coded independently by the first and second authors, with disagreements resolved by discussion. The full list of variables coded is available in the supplementary materials (Cristia et al., 2019a). The most relevant to the present analyses are the following.

We coded number of children included and their demographic characteristics. These demographics included mean age, age range, native language and whether they belong to a special population (e.g. typically developing, at risk for Autism or other developmental disorder, bilingual, twin, or a mix²). Based on this description, we classified a sample as **matching** the

² Studies labeled as “mix” were the following: Alphen, Meester, and Dirks (2017) included children at familial risk of dyslexia and controls; Bredin-Oja, Fielding, and Fleming (2018) had

LENATM training sample (if children were typically-developing 42 months old or less, learners of North American English, with no specific bias for a given socioeconomic status group), or not.

In addition, we coded the size of the samples presented to the annotators. Specifically, we coded whether the annotator was presented with **continuous** samples (i.e., a long section of audio), or with **segments** (i.e., a short stretch that LENA had classified as a specific talker type); when given segments, annotators sometimes were able to inspect a broader context, and at times were able to adjust the LENA-generated boundaries.

We also noted the type of data selection. We coded whether the recordings presented to annotators were chosen based on **high** volubility (CTC, CVC, or AWC), were based on the **algorithm** in some other way (using information from the LENA segmentation algorithm), or whether it was **random**, chosen independently of the LENA segmentation algorithm.³

two children with autism diagnoses, 2 with Down syndrome, 1 with a chromosomal deletion, and 1 with a developmental delay; Lehet, Arjmandi, Dilley, and Houston (2019) included children with a range of hearing statuses, including children with normal hearing, children with hearing aids, and children with cochlear implants; Merz, Maskus, Melvin, He, and Noble (2019) included participants with a range of SES and maternal education statuses; Weisleder and Fernald (2013) included children from low-SES homes who were also bilingual; and D. Xu et al. (2009) included children with a range of SES statuses.

³ Papers that were coded as random for data selection are as follows: in Bredin-Oja et al. (2018) all child vocalizations with clear vocal fold vibration that were not vegetative sounds were transcribed; in Bulgarelli and Bergelson (2019) all utterances containing concrete nouns were

Additionally, total quantity of the audio data annotated was coded, and this at two levels: sample duration in minutes (how long were continuous segments presented to annotators) and total duration (total duration in minutes, collapsing across children and samples)

Some of our analyses below rely on error rates, which are estimates of how close LENA counts are to human counts. When relative error rate (RER) was reported (N=8), we noted how it was implemented. In 8 further cases, RER was not reported but it could be calculated from information provided in the text. In the latter case, we estimated relative error rate as the system estimate minus the human estimate divided by the latter, and multiplied by 100 to have a percentage. Notice that this number is positive when the LENA over-estimates (reports a higher number than the human) and negative when it under-estimates.

Risk assessment at the level of papers was done by assessing whether the authors acknowledged the LENATM Foundation as their affiliation or as a funding source. If any of the authors had such an affiliation, or acknowledged the Foundation, then the paper as a whole was

transcribed; Cristia et al. (2019b) used random or periodic sampling of 1-2 minutes; in Elo (2016) the recordings were transcribed in full; in Jones et al. (2019) a human transcriber started at the beginning of the recording until they encountered a child utterance, at which point they coded 5 full minutes, then skipped 10 minutes and repeated this process; Mccauley, Esposito, and Cook (2011) selected three random 5 minute segments; Oetting, Hartfield, and Pruitt (2009) used the full recording; Schwarz et al. (2017) selected at random; Soderstrom and Franz (2016) transcribed 15 minutes starting 1 hour into the recording; and D. Xu et al. (2009) selected two participants with different SES statuses.

tagged as being at risk. Regarding other risks at the individual level (such as outcome selection or other forms of bias), we systematically coded methodological characteristics that may affect validation results as noted above; for example, the way in which data are selected may involve the LENATM algorithm (in which case generalization to the whole data set may be compromised). We present these methodological variables in context in the Results section. Risk assessment for the whole body of literature was not possible.

Some of the papers (e.g. Bergelson et al., 2018; Bulgarelli & Bergelson, 2019) provided links to their data, allowing us to calculate relevant validation metrics to include here using scripts that can be found in our supplementary materials (Cristia et al., 2019a). Finally, as we will see below, some papers did not provide any metrics that could be integrated quantitatively (Bredin-Oja et al., 2018), but they otherwise match our inclusion criteria and thus are included in qualitative analyses.

Results

Qualitative integration

We found a total of 28 papers reporting 33 (not mutually independent) validation studies, see Table 1. Since only 5 studies involved one or more individuals with a LENATM affiliation, we do not discuss potential bias due to a conflict of interest further. A total of 12 studies from 10 papers have been published in peer-reviewed journals as validation studies, and thus have been (thoroughly) documented and evaluated as such (Bredin-Oja et al., 2018; Bulgarelli & Bergelson, 2019; Busch, Sangen, Vanpoucke, & Wieringen, 2018; Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2016; Ganek & Eriks-Brophy, 2018; Gilkerson et al., 2015; Jones et al., 2019; Oetting et al., 2009; Orena, Byers-Heinlein, & Polka, 2019; VanDam & Silbert, 2016). The

remaining 21 studies from 18 papers are reported in preprints (Berends, 2015; Cristia et al., 2019b; Lehet et al., 2019), theses (Elo, 2016), conference proceedings, posters, or talks (Alphen et al., 2017; Mccauley et al., 2011; Schwarz et al., 2017; Soderstrom & Franz, 2016), white papers (D. Xu et al., 2009), as secondary or preliminary methodological information in the service of a separate research question (Bergelson et al., 2018; Burgess, Audet, & Harjusola-Webb, 2013; Caskey, Stephens, Tucker, & Vohr, 2014; D'Apice, Latham, & Strumm, 2019; Ko, Seidl, Cristia, Reimchen, & Soderstrom, 2016; Merz et al., 2019; Pae et al., 2016), or in an appendix or supplementary material (Seidl et al., 2018; Weisleder & Fernald, 2013). We include this 'grey literature' for thoroughness but encourage any interested readers to filter the data and rerun the analyses with our openly provided code and data as they see fit.

– PLEASE INSERT TABLE 1 HERE –

Studies tended to have small sample sizes, with a median N of 11.50 children ($Range = 1-107$, $M = 22$, Total: 689). A majority ($N = 18$) focused on North American English (NAE); for the other 15, children were learning UK English, US Spanish, Dutch, Finnish, Swedish, France French, Canadian French, Korean, Mandarin and Shanghai Chinese, Tsimane', or Vietnamese. Very few ($N = 8$) reported on children who matched the LENATM's training data sample. The other 25 came from populations that differed from the training sample, including children who were: diagnosed with or at risk for Autism Spectrum Disorder; of markedly low socio-economic status; preterm; twins; particularly high in language skills; bilingual; at risk for developmental delays; or a mixture of these groups. Children's age varied across studies ($M = 35$ months, $Median = 26$ months, $Range = 0-192$ months). Within studies, children's age ranges varied between 0 and 144 months ($M = 26$ months, $Median = 20$ months).

Not all studies reported their process regarding data selection and human annotations. For those that did, 13 studies used a random clip selection algorithm or looked at a whole recording; 6 extracted sections based on high Child Vocalization Count, Conversational Turn Count, and/or Adult Word Count; and a further 10 were otherwise algorithm-dependent. While 14 of the studies presented full clips (e.g., 5 continuous minutes) or whole recordings to annotators, 9 presented sections that had been segmented by the LENATM algorithm. In either case, LENATM annotations may or may not have been visible (and thus bias) the human annotators; and annotators may have been able to resegment them. Unfortunately, not all studies described this aspect of their annotation procedure clearly. The total cumulated duration of annotated data in minutes varied massively across studies (*Range* 25-4200, *Mean* = 809, *Median* = 600).

We now turn to a qualitative overview of validation results. One class of reports pertains to the level of agreement on the labels ascribed to a stretch of audio signal. Two measures are most often reported for this, recall and precision. Recall quantifies accuracy using human tags as the denominator, i.e., $(\text{true positives})/(\text{true positives} + \text{false negatives})$; put otherwise, it answers the question: out of all the vocalizations by talker X, how many got the label X. Precision quantifies accuracy using the system as the denominator, i.e. $(\text{true positives})/(\text{true positives} + \text{false positives})$; put otherwise, it answers the question: out of all the vocalizations labeled as being talker X by the system, how many were really X according to the human annotator. Ideally, such accuracy metrics are reported for each class separately, together with the frequency of that class: Key Child (CHN), Other Child (CXN), Female Adult (FAN), Male Adult (MAN), and none of the above.

However, few of the 33 studies reported even partial recall or precision metrics, and often not at the speech class level of granularity: at most 10/33 report precision/recall on a subset of

labels (e.g. CHN versus CXN, but not FAN, MAN). Instead or in addition, some (8/33) report recall and/or precision for broader classifications (e.g., adult vs. child, collapsing between FAN and MAN on the one hand, and CXN and CHN on the other). Most often than not, non-speech and “far” categories are not discussed at all, and thus it is unclear how any confusions with Silence, Overlap, etc. were handled. A similarly sparse picture appears when we inspect the prevalence of AWC (13 correlations and 14 RERs reported), CVC (5 correlations and 6 RERs reported), and CTC (6 correlations and 5 RERs reported).

We wondered whether this may paint a more dire picture than needed because perhaps the authors did not code their data in a way that would allow them to estimate LENATM accuracy for talkers, or one or more of the other metrics. This case is hard to make for talker identification, which is fundamental to any other validation task: If the authors intend to validate CVC then they need to decide when the key child is speaking and when she is not, thus incidentally producing data for this particular talker category; if they want to validate AWC, they do the same for the adult categories; and if they seek to validate CTC, they should be incidentally producing talker identification data for both children and adult categories. Nonetheless, one can design an annotation scheme where this information is not produced incidentally - for instance, if the annotator has to decide on the fly whether there was a child vocalization or not and then only writes down the total number for the clip. Assuming this worst case scenario, the first author revisited the description of the annotation, human reliability, and results reported, to judge what the authors could have calculated LENATM validity on (regardless of whether they did or did not ultimately report validity results). We found that, in general, validation study reports are less informative than they could be. Specifically, 21 studies could have calculated talker identification accuracy but only 12 reported some recall data and 13 did so for precision. 17 could have

reported on CVC accuracy but only 7 reported a correlation or relative error rate. And 16 could have reported on CTC accuracy, with only 7 providing information on a correlation or a relative error rate. The exception concerned AWC, with all studies evaluating AWC reporting a correlation and/or an error rate.

Moreover, there is considerable variability in methods used. The best example is probably relative error rates (RER), which can be calculated in many ways, for instance by summing counts across all clips and participants; or by calculating relative errors for each clip and participant separately and then averaging this. By and large, if a system is unbiased, the former will produce lower error rates than the latter. Moreover, the interpretation is not the same; if one calculates an error rate from the sum of words transcribed over a whole recording day, the generalization will be across recording days, whereas if the error rates are calculated in smaller time units, then this error rate will generalize more readily to other small time units than to the whole day. Out of the 16 studies in which some RER could be derived (collapsing across AWC, CVC, CTC), we calculated 8 based on reported mean counts, and the remaining 8 were reported on in 4 papers. Two papers did not explain how they had estimated the error rate (although in one case it seems likely that it is the mean of relative error rates calculated separately for each clip and child). One study reported RER separating each of three full days of audio, whereas another also analyzed two full days of audio, but first calculated RER per hour, and then applied the mean to this. In the discussion, we suggest that data sharing may be the best solution to the problem of the blooming diversity in calculation methods.

Quantitative analyses: Central tendencies

As mentioned in the qualitative overview, very few studies provided data at the ideal level of granularity. In light of this sparsity, we made the analytic decision to calculate a global recall

and precision metric, defined as averages of whatever recall/precision were reported within a study. This allowed us to consider 12 studies (36% of all studies) for recall, and 13 (39%) for precision, i.e., fewer than half of studies could contribute data to this analysis even with this lax definition. Problematically, authors may only report the categories for which they obtained relatively high recall and/or precision. If so, this could introduce outcome selection bias into our results. With that caveat in mind, we report that overall recall and precision estimates are quite high (**recall**: N=12, mean = 59%; N=12, weighted mean = 62%, where weights are based on the total cumulated duration of annotated data; **precision**: N=13, mean = 68%; N=11, weighted mean = 64%).

Regarding LENATM's numeric output (AWC, CVC, CTC), evaluations are based on correlations (r) and/or some form of relative error rate (RER). As a reminder, most of these RERs come from taking the sum or average count for the system minus the same for human annotators divided by the latter (multiplied by 100 to have a percentage), such that a positive number indicates that the LENA over-estimates and negative that it under-estimates. The Pearson r for AWC was very high (N=13, mean = 0.79; N=13, weighted mean = 0.79), with relatively low RERs suggesting a slight tendency to over-estimate AWC (N = 14, mean = 13.76%; N = 13, weighted mean = 4.48%).

For CVC, the Pearson r was also quite high (N = 5, mean = 0.77; N = 5, weighted mean = 0.74), but RERs were numerically larger and indicated a underestimate CVC (N = 6, mean = -24.17%; N = 6, weighted mean = -40.48%, which would mean CVCs are nearly halved). For CTC, the Pearson r was quite low (N = 6, mean = 0.36; N = 6, weighted mean = 0.36) and the RERs also indicated a rather strong underestimation tendency (N = 5, mean = -34%; N = 4, weighted mean = -50%).

Quantitative analyses: Participant and Methodological Moderators

We next intended to carry out a meta-regression with moderators in order to assess whether participant-specific or methodological factors impacted recall, precision, or AWC estimate accuracy. As a first approach, we provide readers with figures summarizing trends in the data. These figures highlight the fact that, with such a small number of points, the variability is large enough that no differences can be detected, with low confidence even about whether line slopes are positive or negative for continuous variables.

Figure 2 displays our three dependent measures as a function of the infants' native language (North American English or other), population sample (matching the sample in the LENA™ training or not), and age (mean and range within each study). Figure 3 depicts our dependent measures split by methodological variants: whether the clips were selected at random or using the LENA™ algorithm (to identify periods with high word counts or in any other way); how long clips were; and how much data were annotated.

–INSERT FIGURE 2 ABOUT HERE –

–INSERT FIGURE 3 ABOUT HERE –

For both sets of graphs, in the top panels the central tendency is depicted via mean and 95% non-parametric bootstrapped confidence intervals; in the bottom panels, a linear fit is plotted in blue, using R's `predict()` function to draw a grey 95% confidence band around it. Notice that confidence intervals are wide and overlap across the relevant variables; the confidence bands are also consistent with positive or negatively sloping lines. This suggests that the level of variability masks any potential differences as a function of participant and methodological variables.

We have decided to omit inferential statistics because of several concerns about their interpretation and validity. First, not all data points are mutually independent, which violates basic assumptions of most tests. Second, as the figures highlight, many of the comparisons would be based on 3-5 data points in each cell, which may lead to false negatives (i.e., if we conclude there is no difference in recall/precision/AWC for a potential moderator but in reality the test lacked power to detect it), and false positives (i.e., if we conclude there is a difference in recall/precision/AWC for a potential moderator but this is due to a chance drawing of extreme values which have an overblown effect given very small sample sizes per cell). We encourage readers who would like to perform analyses—which we feel are likely premature—to download the data, available from our online Supplementary Materials (Cristia et al., 2019a).

Discussion

The number of publications using LENATM has grown steadily in the preceding decade, with many of these studies including statements that the method has been successfully validated for the American English early childhood sample it has been developed for, as well as for other samples. Indeed, a handful of validation studies are scattered in the literature, sometimes as the primary goal of a paper, and sometimes as an appendix or footnote. The present paper sought to systematically review and evaluate those studies in order to inform the field on the strength of the evidence for appropriate validation of LENATM metrics. Additionally, we used quantitative integration to estimate overall accuracy of the LENATM system, and we intended to explore the possibility that its accuracy is moderated by participant and/or methodological factors.

We found a sizable number of studies purporting to report how well the LENATM system fared relative to human annotators in terms of classification precision and recall, and/or derived

Adult Word Counts, Conversational Turn Counts, and Child Vocalization Counts. There were 33 studies appearing in 28 different papers. A systematic review of these data revealed that few studies fully reported all relevant classification information and/or derived counts (with as few as 5 non-independent points for Child Vocalization Counts). Moreover, few report the full information necessary to assess how exactly the validation took place. For instance, full confusion matrices are virtually never provided, making it impossible to know exactly what was analyzed. That is, authors may have altogether excluded stretches labeled as Silence, Overlap, and all “far” labels by LENATM, which therefore means that any confusion of the speech categories with these other categories is not counted against the algorithm’s accuracy.

To a certain extent, lack of information is likely a side effect of these validations being done in service of a specific research question and thus reported on very briefly, sometimes not in the main manuscript at all. In fact, only 36% of the papers had been peer-reviewed as a validation study, and even these did not report all information needed to interpret results in the context of day-long recordings at large. To make this more concrete, consider studies reporting precision and recall. Only 4 were published in a peer-reviewed journal as validation studies (in Table 1: Bulgarelli & Bergelson, 2019; Gilkerson et al., 2015; Jones et al., 2019; VanDam & Silbert, 2016). To pick on our own work, Bulgarelli & Bergelson (2019), only segments containing concrete nouns were tagged for speaker identity, leading to precision and recall estimates that are hard to generalize to day-long data at large, particularly for the key children (who were 6-7 months of age). Specifically, since infants this age are not saying any words, any section that was tagged as them by the LENATM system could only count as a false positive; and no utterances by the child that the LENATM system was successful in finding would count in its favor. A more common issue is collapsing across LENATM categories (Gilkerson et al., 2015; VanDam &

Silbert, 2016), reporting on accuracy of the algorithm when distinguishing children (collapsing between Key Child and Other Child) from adults (collapsing between Female Adult and Male Adult). By collapsing in this way, any confusion within the collapsed categories is not penalized, again inflating accuracy estimations. Moreover, some LENATM end users may not realize that this level of accuracy cannot be expected in samples where there are many other children. A third issue we would like to point out relates to the segmentation, that is, how stretches of the acoustic signal are labeled. It is very common for annotators to hear LENATM segments, in (Bulgarelli & Bergelson, 2019; Gilkerson et al., 2015; Jones et al., 2019) or out (VanDam & Silbert, 2016) of context. In the latter case, any interstitial silences and even talker overlap may be ignored by the coders. In the former, it may be easier for annotators to notice that one subsection of the segment is actually the end of the utterance by another speaker, but it is not always the case that annotators can actually signal this and correct for it. For instance, Gilkerson et al. (2015) state in the methods section that coders could correct the LENATM segmentation, but in the results they do not mention how often this was necessary. In their Study 1, Jones et al. (2019) did report that re-segmentation was rarely needed.

Moreover, since it is not standard to report all methodological choices and validation results, it is possible that authors (present company included) may have reported only “relevant” analyses, or determined whether accuracy was “good enough” for a given purpose, especially when the validation served as a stepping-stone to a further question of interest (e.g., Bergelson et al., 2018; Seidl et al., 2018). As in other fields, bias may be deepened by conflicts of interest involving authors and funders; but here, bias could also seep in simply if authors need to justify to reviewers their use of this technology, making it advantageous or more streamlined to combine categories or skip troublesome regions. This falls within the umbrella of “outcome selection bias”

in systematic review terms. To be clear, we do not claim any authors are acting maliciously or deceptively, simply that by underreporting validation information, it becomes incredibly difficult to evaluate the evidence base underlying the accuracy of LENATM output.

We interpret this evidence as a clear signal that authors publishing validation studies have adopted underinformative habits, as elucidated by our critical consideration of the literature here. We hope this systematic review will lead to a course correction in future work.

With these general considerations in mind, we turn with tempered enthusiasm to a discussion of the otherwise good results we saw in our quantitative integration of reported data: recall and precision higher than 58.80% based on up to 13 non-independent studies; strong correlations for AWC (mean of $r = 0.79$, on $N = 13$, and a mean RER of 13.76, on $N = 14$); and a similarly high correlation for CVC (mean of $r = 0.77$, on $N = 5$, with a mean RER = -24.17, on $N = 6$). The exception to this general trend towards good performance was CTC, with a mean correlation of $r = 0.36$, on $N = 6$, and a mean RER = -34.20, on $N = 5$. Given the importance that current theories and descriptions give to LENATM measures of conversational turns (e.g., Merz et al., 2019; Romeo et al., 2018), it is important to collect more validation data on conversational turn counts, if possible using unbiased data selection methods (i.e., not using the LENATM algorithm segmentation as a starting point) and larger samples than currently used (the mean number of infants in CTC validation studies is 20.60, each represented by 10.28 minutes of audio). The data on relative error rates (RER) was extremely variable, in part due to variability in implementation.

One important limitation of the present analysis is that recall and precision could only be inspected at a gross level, by averaging across any accuracies that were reported in each study.

Thus, while the LENATM system could potentially give us insight on many important aspects of naturalistic input, we are not yet in a position to make firm conclusions about its accuracy. For instance, it is extremely relevant to separate input produced by the key child (wearing the recorder) versus other children in the environment for many basic science and clinical applications (including analyzing speech in daycares or families with more than one child). However, only 8 non-independent studies report recall and precision on the Key Child label. Moreover, even though some of them omit the confusions between the Key Child category and non-speech categories (such as Silence and Overlap), which likely inflates accuracy, current results show a mean recall of 56% and a mean precision of 55.88%. Similarly, categories like Silence, Overlap, Electronics, and subcategories within the key child (speech versus vegetative sounds and crying) are so rarely included in validation efforts as to preclude their inclusion in this systematic review altogether.

The results above can also help guide our field's approach in uncovering the stability of LENA accuracy across participant moderators. Specifically, the missingness we note leads to several clear recommendations of specific areas in which validation data are sparse. In examining accuracy as a function of whether the sample matches the sample LENATM algorithms were trained on, we found that more work is needed particularly in terms of language. There were very few data points on labels' recall and precision for samples that were not exposed to north-American English (non-NAE; for example for recall there were 4 non-NAE, versus 8 NAE). The opposite occurs for AWC, with only 3 studies on NAE as opposed to 10 studies on non-NAE. We therefore recommend more studies evaluating non-NAE samples for recall and precision of speech classes, and more NAE samples evaluating AWC accuracy. Regarding mean age, there is a trend for lower recall for data sampled from older children, with no such tendency for precision

or AWC. Most data points, however, cluster in the first 2 years of life. Therefore, we recommend continued efforts to assess LENA accuracy with older children.

For authors who are inspired by the wide space of still-important validation to be done, and for those who are already conducting validation as part of their research, we would like to recommend that they follow standardized procedures to facilitate comparisons across papers and reduce the impact of algorithm and researcher bias. Some recommendations have been laid out in other work (Cristia et al., 2019b), including (1) a preference for coding continuous clips rather than individual segments, (2) sampling semi-randomly to ensure that selection is not biased by the algorithm, and (3) using the DARCLE Annotation Scheme to increase comparability of the annotation (Casillas et al., 2017). Regardless of how data are sampled and annotated, we strongly encourage authors of validation work to make the raw validation data publicly available. Data containing automated and manual categorical and numerical information are intrinsically deidentified, and thus bypass most of the usual privacy and ethics-based concerns that apply to the audio data itself. The simplest way to share them is through websites like osf.io, which in a matter of minutes allows you to create a public-facing project (with the option of private components as well), containing a Digital Object Identifier (DOI) that can then be readily cited by others. This provides the original authors with ‘credit’ for their validation data, which is separate from the specific research questions addressed in their original empirical reports. In addition to being beneficial to authors, sharing such data would allow others to re-visit them and calculate statistics and metrics not considered by the original authors. For instance, as mentioned above, validating CTC requires identifying adults separately from children, and thus data originally created to validate CTC can be re-used to validate the system’s categorization into child and adult voices. The current systematic review already benefitted from being able to

calculate metrics not reported in original papers thanks to authors' posting their data and code (Bergelson et al., 2018; Bulgarelli & Bergelson, 2019). Moreover, such data sharing is also useful for measures that can be computed in different ways, such as error rates. Shared data would allow meta-analysts to explore various implementations of such error rates with no effort or time required from the original authors.

Finally, as reviewers and editors, we should give papers reporting nothing but a validation effort our full attention: If the field continues to use the LENATM technology for both intervention and basic science (Sosa, 2016; e.g., D. L. Suskind et al., 2016), it becomes crucial to independently establish its validity. In addition, acceptance should be based on the validation approach and not on the ensuing results, to ensure that studies showing low validity with high quality methods have a chance of entering the literature. We hope in particular that this systematic review serves as an opportunity for those using LENATM to take our freely and openly available data, code book, and scripts to make sure their validity experiments using the automated LENATM measures or other analyses of naturalistic recordings can be integrated into meta- or mega-analyses. In the future, another systematic review that is pre-registered in an available database would then be in a better position to present the most unbiased estimation of the LENATM algorithms and other similar systems that may emerge.

Acknowledgements

We are grateful to members of the DARCLE Consortium for comments on previous versions of this work.

References

- Adams, K. A., Marchman, V. A., Loi, E. C., Ashland, M. D., Fernald, A., & Feldman, H. M. (2018). Caregiver talk and medical risk as predictors of language outcomes in full term and preterm toddlers. *Child Development, 89*(5), 1674–1690.
<https://doi.org/https://doi.org/10.1111/cdev.12818>
- Alphen, P. van, Meester, M., & Dirks, E. (2017). LENA onder de loep.
- Berends, C. (2015). The LENA System in Parent-Child Interaction in Dutch Preschool Children with Language Delay.
- Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2018). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, (June 2018), 1–12.
<https://doi.org/https://doi.org/10.1111/desc.12724>
- Bredin-Oja, S. L., Fielding, H., & Fleming, K. K. (2018). Clinician vs. Machine: Estimating Vocalizations Rates in Young Children With Developmental Disorders. *American Journal of Speech-Language Pathology, 27*(3), 1066–1072.
https://doi.org/http://dx.doi.org/10.1044/2018_AJSLP-17-0016
- Bulgarelli, F., & Bergelson, E. (2019). Look who’s talking: A comparison of automated and human-generated speaker tags in naturalistic daylong recordings. *Behavior Research Methods*. <https://doi.org/https://doi.org/10.3758/s13428-019-01265-7>
- Burgess, S., Audet, L., & Harjusola-Webb, S. (2013). Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with ASD.

Journal of Communication Disorders, 46(5-6), 428–439.

<https://doi.org/https://doi.org/10.1016/j.jcomdis.2013.09.003>

Busch, T., Sangen, A., Vanpoucke, F., & Wieringen, A. van. (2018). Correlation and agreement between Language ENvironment Analysis (LENATM) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, 50(5), 1921–1932.

<https://doi.org/https://doi.org/10.3758/s13428-017-0960-0>

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

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017). A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of childrens language environments. In *Interspeech 2017* (pp. 2098–2102).

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2014). Adult Talk in the NICU With Preterm Infants and Developmental Outcomes:e578-e584. *Pediatrics*, 133.

<https://doi.org/10.1542/peds.2013-0104>

Cristia, A., Bulgarelli, F., & Bergelson, E. (2019a). Accuracy of the Language Environment Analysis System: A systematic review - Supplemental materials. Retrieved from

<https://osf.io/4nhms/>

- Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C. F., Rasanen, O., ... Bergelson, E. (2019b). A thorough evaluation of the Language Environment Analysis (LENA) System. <https://doi.org/10.1017/CBO9781107415324.004>
- D'Apice, K., Latham, R. M., & Strumm, S. von. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Developmental Psychology*. <https://doi.org/https://doi.org/10.1016/j.pec.2013.10.008>
- Elo, H. (2016). *Acquiring language as a twin*. Tampere, Finland: Tampere University Press.
- Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language ENvironment Analysis (LENA) Conversational Turn Counts in Vietnamese. *Communication Disorders Quarterly*, 39(2), 371–380. <https://doi.org/https://doi.org/10.1177/1525740117705094>
- Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). Transcriptional analyses of the LENA natural language corpus LTR-06-2. *Boulder, CO: LENA ...*, (September 2008), 1–11.
- Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., ... Toppings, K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech Language and Hearing Research*, 85(April), 3000. <https://doi.org/https://doi.org/10.1044/2015>
- 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/https://doi.org/10.1177/1525740110367826>

- Jones, R. M., Plesa Skwerer, D., Pawar, R., Hamo, A., Carberry, C., Ajodan, E. L., ... Tager-Flusberg, H. (2019). How effective is LENA in detecting speech vocalizations and language produced by children and adolescents with ASD in different contexts? *Autism Research*, 1–8. <https://doi.org/https://doi.org/10.1002/aur.2071>
- Ko, E.-S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. *Journal of Child Language*, 43(2), 284–309. <https://doi.org/https://doi.org/10.1017/S0305000915000203>
- Lehet, M., Arjmandi, M. K., Dilley, L. C., & Houston, D. (2019). Accuracy of the Language ENvironment Analysis (LENA) system for quantifying adult speech experienced by infants in naturalistic settings.
- Mccauley, A., Esposito, M., & Cook, M. (2011). Language Environment Analysis of Preschoolers with Autism: Validity and Application. *Poster*.
- Merz, E. C., Maskus, E. A., Melvin, S. A., He, X., & Noble, K. G. (2019). Socioeconomic disparities in language input are associated with children’s language-related brain structure and reading skills. *Child Development*.
<https://doi.org/https://doi.org/10.1111/cdev.13239>
- Oetting, J. B., Hartfield, L. R., & Pruitt, S. L. (2009). Exploring LENA as a tool for researchers and clinicians. *ASHA Leader*, 14(6), 20.
<https://doi.org/https://doi.org/10.1044/leader.FTR3.14062009.20>
- Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Cutler, E. A. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language

delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30), 13354–13359. <https://doi.org/https://doi.org/10.1073/pnas.1003882107>

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the Language Environment Analysis Recording System in Analyzing French–English Bilingual Speech. *Journal of Speech, Language, and Hearing Research*, 62(7), 2491–2500.
https://doi.org/https://doi.org/10.1044/2019_jslhr-l-18-0342

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J. A., Ma, L., & Topping, K. (2016). Effects of feedback on parent-child language with infants and toddlers in Korea. *First Language*, 36(6), 549–569. <https://doi.org/https://doi.org/10.1177/0142723716649273>

Richards, J. A., Gilkerson, J., Paul, T. D., & Xu, D. (2008). The LENATM automatic vocalization assessment LTR-08-1. *Boulder, CO: LENA ...*, (September 2008), 1–12.

Romeo, R. R., Segaran, J., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., ... Gabrieli, J. D. E. (2018). Language exposure relates to structural neural connectivity in childhood. *The Journal of Neuroscience*, 38(36), 7870–7877.
<https://doi.org/https://doi.org/10.1523/JNEUROSCI.0484-18.2018>

Schwarz, I.-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., & Marklund, E. (2017). The LENA TM system applied to Swedish: Reliability of the Adult Word Count estimate. *Interspeech*, 2088–2092.

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. (2018). Infant–Mother acoustic–Prosodic alignment and developmental risk. *Journal of Speech, Language, and Hearing Research*, 61(6), 1369–1380.

Soderstrom, M., & Franz, W. (2016). Comparing human- and machine-annotated language input across childcare settings. *Talk Presented at ICIS 2016*.

Sosa, A. V. (2016). Association of the Type of Toy Used During Play With the Quantity and Quality of Parent-Infant Communication. *JAMA Pediatrics*, *170*(2), 132.

<https://doi.org/https://doi.org/10.1001/jamapediatrics.2015.3753>

Suskind, D. L., Leffel, K. R., Graf, E., Hernandez, M. W., Gunderson, E. A., Sapolich, S. G., ...

Levine, S. C. (2016). A parent-directed language intervention for children of low socioeconomic status: a randomized controlled pilot study. *Journal of Child Language*,

43(02), 366–406. <https://doi.org/https://doi.org/10.1017/S0305000915000033>

Suskind, D. L., Leffel, K. R., Hernandez, M. W., Sapolich, S. G., Suskind, E., Kirkham, E., &

Meehan, P. (2013). An exploratory study of “quantitative linguistic feedback”.

Communication Disorders Quarterly, *34*(4), 199–209.

<https://doi.org/https://doi.org/10.1177/1525740112473146>

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), e12456. <https://doi.org/https://doi.org/10.1111/desc.12456>

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE*, *11*(8), 1–8.

<https://doi.org/https://doi.org/10.1371/journal.pone.0160588>

Weisleder, A., & Fernald, A. (2013). Talking to children matters. *Psychological Science*, *24*(11),

2143–2152. <https://doi.org/https://doi.org/10.1177/0956797613488145>

- Wood, C., Diehm, E. A., & Callender, M. F. (2016). An investigation of language environment analysis measures for spanish–English bilingual preschoolers from migrant low-socioeconomic-status backgrounds. *Language, Speech, and Hearing Services in Schools*, 47(2), 123–134. https://doi.org/https://doi.org/10.1044/2015_LSHSS-14-0115
- Xu, D., Yapanel, U., & Gray, S. (2009). Reliability of the LENA Language Environment Analysis system in young children’s natural home environment. *LENA Foundation Technical Report*, (February), 1–16.

PRISMA 2009 Flow Diagram

Figure 1: PRISMA Flowchart

Figure 2: Outcomes by Participant Moderators. Top left panel: Infant language (within panel: left side shows NAE = N(orth) American English while right depicts other languages). Top right panel: Match of infant population to LENA training sample (within panel: matching samples on the right, mis-matching on the left). Bottom panels: infant mean age (bottom left) and infant age range (bottom right). Each point indicates one study, numbers indicate study identity (see Table 1). Filled square points indicate authors affiliated with LENA. Y axes indicates the scale for the variable indicated in the panel title (e.g. precision). N.B. axes values vary since different studies may be included across panels, depending on what papers reported. Red lines indicate bootstrapped confidence intervals (CI); grey bands in bottom panel indicate 95% CIs from a linear fit to the data. See text for details and interpretive caveats.

Figure 3: Outcomes by Methodological Moderators. Top left panel: Segment selection by algo(rithm) (left) versus randomly (right). Top right panel: Segment size (continuous (left) versus single-segment (right)). Bottom left panel: Duration of individual samples. Bottom right panel: Total cumulative annotated data. Each point indicates one study, numbers indicate study identity (see Table 1). Filled square points indicate authors affiliated with LENA. Y axes indicates the scale for the variable indicated in the panel title (e.g. precision). N.B. axes values vary since different studies may be included across panels, depending on what papers reported. Red lines indicate bootstrapped confidence intervals (CI); grey bands in bottom panel indicate 95% CIs from a linear fit to the data. See text for details and interpretive caveats.

Table 1: Papers and studies included in the systematic review: Precision and recall. Note that not all studies contribute data for quantitative integration. Regardless of how many authors a publication has, papers are identified by the first author and year of publication. Papers with * denote LENA affiliation, those with + were published as evaluations in peer reviewed journals. Recall and precision are provided in percentage points. “Included in recall” lists the sources of recall values included in that average recall; “included in precision” the same for precision: Adult = adult categories collapsed, Child = child categories collapsed, CHN = target child, CXN = other children, FA = female adult, MA = male adult. All numeric predictors have been rounded for this display.

id	paper	recall	included in recall	precision	included in precision
1	Alphen 2017				
2	Berends 2016				
3	Bergelson 2018			84	FA, MA
4	Bredin-Oja 2018+				
5	Bulgarelli 2019+	62	Adult, Child, CXN, FA, MA	63	Adult, Child, CHN, CXN, FA, MA
6	Burgess 2013				
7	Busch 2018+				
8	Canault 2016+				
9	Caskey 2014				
10	Cristia 2019	28	Adult, Child, CHN, CXN, FA, MA	41	Adult, Child, CHN, CXN, FA, MA
11	D'Apice2019				
12.1	Elo 2016	86	CHN, CXN, FA, MA	81	CHN, CXN, FA, MA
12.2	Elo 2016	86	CHN, CXN, FA, MA	90	CHN, CXN, FA, MA
13	Ganek 2018+				
14	Gilkerson 2015*+	80	Adult, Child	46	Adult, Child, CHN, CXN, FA, MA
15.1	Jones 2019+	56	CHN		
15.2	Jones 2019+	11	CHN		
15.3	Jones 2019+	46	CHN	61	Adult, CHN
16	Ko 2015			84	CHN, FA
17	Lehet 2019	62	Adult, Child, FA, MA	64	Adult, Child, FA, MA
18	McCauley 2011	64	Adult, Child, CHN, CXN		

19	Merz 2019		
20	Oetting 2009+		
21	Orena 2019+		
22	Pae 2016*		
23	Schwartz 2017		
24	Seidl 2018		72 CHN, FA
25	Soderstrom 2016	45 Adult, Child, CHN, CXN, FA, MA	62 Adult, Child, CHN, CXN, FA, MA
26	VanDam 2016+		71 Adult, Child, FA, MA
27	Weisleder 2013		
28.1	Xu 2009*	79 Adult, Child	69 Adult, Child
28.2	Xu 2009*		
28.3	Xu 2009*		

Table 2: Papers and studies included in the systematic review: AWC, CVC, CTC. Note that not all studies contribute data for quantitative integration. Regardless of how many authors a publication has, papers are identified by the first author and year of publication. Papers with * denote LENA affiliation, those with + were published as evaluations in peer reviewed journals. AWC stands for adult word counts; CVC child vocalization counts, CTC for conversational turn counts. R for correlations between human and LENA counts, RER for relative error rates (in percentage points). All numeric predictors have been rounded for this display.

id	paper	AWC_R	AWC_RER	CVC_R	CVC_RER	CTC_R	CTC_RER
1	Alphen 2017		18				
2	Berends 2016						-90
3	Bergelson 2018						
4	Bredin-Oja 2018+						
5	Bulgarelli 2019+						
6	Burgess 2013		27				
7	Busch 2018+	0.87	-20	0.77	2	0.52	-64
8	Canault 2016+	0.64	-33	0.71	-66		
9	Caskey 2014	0.93					
10	Cristia 2019	0.75	55	0.8	-20	0.59	-67
11	D'Apice2019	0.79	12				
12.1	Elo 2016		49		-10		
12.2	Elo 2016		67		-6		
13	Ganek 2018+					0.7	43
14	Gilkerson 2015*+	0.73	9			0.22	7
15.1	Jones 2019+						
15.2	Jones 2019+						
15.3	Jones 2019+						
16	Ko 2015						
17	Lehet 2019						
18	McCauley 2011			0.81	-45		
19	Merz 2019			0.74			
20	Oetting 2009+	0.85	18			0.14	

21	Orena 2019+	0.77	20	
22	Pae 2016*	0.72		-0.03
23	Schwartz 2017	0.67		
24	Seidl 2018			
25	Soderstrom 2016	0.82		
26	VanDam 2016+			
27	Weisleder 2013	0.8		
28.1	Xu 2009*	0.92	-2	
28.2	Xu 2009*		-0.4	
28.3	Xu 2009*		-27	

Table 3: Papers and studies included in the systematic review: Demographic and methodological characteristics. Regardless of how many authors a publication has, papers are identified by the first author and year of publication. Papers with * denote LENA affiliation, those with + were published as evaluations in peer reviewed journals. Language stands for native language of participants, NAE = North American English, UKE = United Kingdom English, Chn = Shanghai Chinese and/or Mandarin Chinese, Sp = Spanish, Fr = French, Vietn = Vietnamese; Tsi = Tsimane'. Total: total duration of annotated samples (in minutes). Size stands for size of the sample provided to the human annotator: segm = segment, cont = continuous. Sel stands for type of selection: random = unrelated to LENA segmentation, high = based on high AWC/CVC/CTC, other = based on LENA segmentation in some other way. Participant stands for characteristics of sample: typ = typically developing and with none of the other characteristics, ASD(risk) = diagnosed with (or at risk for) Autism Spectrum Disorder, highL = high language, lowSES = family with low socio-economic status, Drisk = at risk for developmental delays, pret = preterm, twin = having a twin, biling = being bilingual, mixed = a mixture of any of the above. Age = mean age of participants in months. Range = age range of participants in months (except for Caskey 2014, where it indicates weeks gestation, wg). N is the number of children included in the sample. Match reflects whether it matches the LENA training sample. All numeric predictors have been rounded for this display.

id	paper	language	total	sel	size	participant	age	rng	N	match
1	Alphen 2017	Dutch				mix	37.9		42	no
2	Berends 2016	Dutch		high		Drisk	36	30-42	14	no
3	Bergelson 2018	NAE		algo_driven	segm	typ.	11.5	5-20	61	yes
4	Bredin-Oja 2018+	NAE	138	random	cont	mix	36	28-46	6	no
5	Bulgarelli 2019+	NAE	1932	random	segm	typ.	6.5	6-7	44	yes
6	Burgess 2013	NAE	465	high	cont	ASD	51	35-67	10	no
7	Busch 2018+	Dutch	240	algo_driven	cont	typ.	42	24-60	5	no
8	Canault 2016+	Fr	3240	high		typ.	25.5	3-48	18	no
9	Caskey 2014	NAE, Sp	25			preterm	0	32-36 wg	5	no
10	Cristia 2019	NAE, UKE, Tsimane	1472	random	cont		16.8	3-58	49	no
11	D'Apice 2019	UKE	320	high	cont	typ.	33.2	24-48	107	no
12.1	Elo 2016	Finnish	698	random	cont	twin	7	0	1	no
12.2	Elo 2016	Finnish	630	random	cont	twin	9	0	1	no

13	Ganek 2018+	Viet	100	algo_driven	cont	mix	30. 5	22-42	10	no
14	Gilkerson 2015*+	Chn	330	high	cont	typ.	12. 1	3-23	22	no
15.1	Jones 2019+	NAE	120	random	cont	ASD	78	60-96	8	no
15.2	Jones 2019+	NAE	105	random	cont	ASD	192	168-216	7	no
15.3	Jones 2019+	NAE	2210		segm	ASD	96	60-204	36	no
16	Ko 2015	NAE	26	algo_driven	segm	typ.	20. 4	12-30	13	yes
17	Lehet 2019	NAE	734	algo_driven	cont	mix	20	4-34	23	no
18	McCauley 2011	NAE	150	random		ASD			5	no
19	Merz 2019	NAE	600	algo_driven		mix	90	61-119	10	no
20	Oetting 2009+	NAE	510	random	cont	lowSES	42	24-60	17	no
21	Orena 2019+	Fr, NAE	945	algo_driven	segm	bilingual	10	10-11	21	no
22	Pae 2016*	Korean	630			typ.	12. 5	4-16		no
23	Schwartz 2017	Swedish	240	random	cont	typ.	30		4	no
24	Seidl 2018	NAE		algo_driven	segm	ASDrisk			10	no
25	Soderstrom 2016	NAE	1305	random	segm	typ.	25	12-38	32	yes
26	VanDam 2016+	NAE	47	algo_driven	segm	typ.	29. 1		26	yes
27	Weisleder 2013	Sp	600	algo_driven		mix	19	0	10	no
28.1	Xu 2009*	NAE	4200	high	segm	mix	19	2-36	70	yes
28.2	Xu 2009*	NAE	720	random		typ.	10	0	1	yes
28.3	Xu 2009*	NAE	720	random		highL	31	0	1	yes